An Efficient Way to Handle the High Dimensional Problem with Fuzzy Association Rule

Indhumathi. V
Department of Computer Science and Engineering M.A.M College of Engineering Siruganur, Trichy,

Karthika. S K
Department of Computer Science and Engineering M.A.M College of Engineering Siruganur, Trichy,

***Abstract*—Prediction is one of the important tasks in data mining process which predicts the future events and unknowns in the form of continuous valued functions. Mining sequential rule in data mining has limitations like very specific rules and similar rules representing the same situation. This creates three major problems such as rating of similar rules can be different, rules are not found because the individual rules are considered uninteresting and too specific rules have less chance for making predictions. In this paper, these issues can be addressed using the idea of mining "Partially Ordered Sequential Rules (POSR), where items in the antecedent and consequent are unordered. Fuzzy logic is also used in this paper to remove the infrequent items. RuleGrowth and TRule Growth algorithm are proposed to mine POSR where TRule Growth algorithm accepts the sliding window constraint to find rules in maximum amount of time which has excellent scalability and provides higher prediction accuracy.**

Keywords*—Prediction; Sequential Rule; Fuzzy logic; Rule Growth; TRule Growth; sliding window constraint

I. INTRODUCTION

The most important research domain in this 21st century is the Data Mining, the process of examining large pre-existing database to generate new information. The goal is to extract only frequents from a large amount of data. Several algorithms have been proposed for this purpose of extraction. Prediction is final goal in data mining to predict something that will happen in future. For example, predicting the future value of a company in a stock market analysis, weather condition, and risk of individual diseases and also behavior of the learners.

Major issues concerned with prediction are speed, scalability, predictive accuracy, goodness of rules. Before starting the prediction the dataset need to be classified first as to make the process of prediction easier. The existing work uses many algorithms such as GSP, SPADE, and PrefixSpan for sequential pattern mining. These algorithms mislead the user to make decisions because the patterns found by these algorithms are only based on the support value. Solution to this problem is to add a measure of confidence but adding this to sequential pattern is not straightforward because it contains multiple items. An alternative approach is the sequential rule (SR) mining has three drawbacks. First, rules may have many variations with different item ordering. Second, the ratings by the algorithms show differences between the rules and their variations. Third, rules have less chance for making prediction because of specific rules.

The proposed work in this paper overcomes these issues by introducing the partially-ordered sequential rules (POSR), where items are unordered. The algorithms RuleGrowth and TRuleGrowth proposed for mining partially-ordered sequential rules find rules by applying constraints. The constraint is to find rules with a sliding window because users often wish to discover rules within the maximum amount of time. To find rules only frequent items are extracted the infrequent items are eliminated using the fuzzy logic technique. Linguistic terms and membership functions are used removing those infrequent items.

This paper is organized as follows: Section 2 presents the related works on sequential rule mining and association rule. Section 3 describes the proposed work. Section 4 concludes the proposed work with the observations.

II. RELATED WORKS

Several algorithms and techniques have been proposed and used for the sequential rule and association rule mining. In [7] Phillippe Fournier Viger et al. proposed an algorithm named TNS (Top k-Non redundant Sequential rule) avoids the generation of redundant sequential rule based on the depth first search strategy which shows excellent performance and scalability. The new data structure named CMAP (Co- occurrence MAP) is suggested by Antonio Gomariz et al. in [10] to store co-occurrence information to overcome the performance bottleneck problem of vertical representation in sequential pattern mining. CMAP is used to prune the candidate generation in three state-of-the-art vertical algorithms namely SPADE, SPAM and GSP.

Ted Gueniche in [11] suggested a framework called SPMF (Sequential Pattern Mining Framework) is a cross platform library in java for discovering patterns. Command line interface and graphical interface that are offered by the sequential pattern mining framework performs fast testing. Algorithms are framed during the source code version and released during the release version are the two versions in SPMF. Predicting the sequential data is research

problem with many applications like stock market prediction and web link recommendation. Instead of using mining sequential rules for large sequences, partially-ordered sequential rule is new approach in [9] by Phillippe Fournier Viger et al. this improves the accuracy of prediction and the matching rate. In [6] Phillippe Fournier Viger et al. proposed an algorithm CMRules scans the sequence database and finds all the association rules by applying the algorithm of association rule mining algorithms like Apriori. Original database also scanned for calculating the sequential support and sequential confidence for each association rule. Finally the association rules are eliminated (SeqSup(r) < minSeqSup and SeqConf(r) < minSeqConf) and return the remaining set of rules. David Lo et al. in [17] focused on recovering rules from the execution traces. Propositional rule mining algorithm is used which enables the mining to be complete by avoiding exponential blowup. The results produced by this approach are efficient and very effective.

A syntactic characterization of a non-redundant sequential rules is investigated and proposed in [16] by David Lo et al. built on the representative patterns. A rule is said to be redundant if the support and confidence values are same. CNR (Compressed set of Non-Redundant) rule mining algorithm is proposed based on the definition of configuration key that improves the runtime and compactness of the rule. Minqing Hu et al. focused on extracting the opinion features from the benefits and drawbacks in [13]. For this purpose a language pattern based approach is proposed and the patterns are generated from the CSR (class sequential rules). The class sequential rule is different from the classic sequential rule where CSR has fixed class showing the implication as X → Y. Jay Ayres et.al, introduced new efficient algorithm in [3] named SPAM (Sequential Pattern Mining). This algorithm uses the depth first strategy and uses vertical bitmap data allowing for simple and efficient counting. The lexicographic tree is traversed for sequences and the pruning methods are used to reduce the search space. The problem of finding rules by relating the patterns to the time series is considered in [5] by Gautham Das et al. Adaptive methods are described based on discretizing the sequence in the database by resembling the methods as vector quantization. Subsequence are first formed by sliding window and based on the similarities these subsequence are clustered. Rakesh Agrawal et al. in [1] presented an efficient template algorithm. The algorithm generates all significant rules between the items which incorporates the buffer management, estimation techniques and pruning techniques. Rules are found with the minimum support and minimum confidence. In [2] two algorithms have been proposed AprioriSome and AprioriAll by Rakesh Agrawal et al. AprioriSome find patterns with minimum number of customers and the sequential pattern found by this is low. AprioriAll algorithm scales up linearly with the maximum number of customers.

Sherri K. Harms et al. in [12] provided a new approach MOWCATL (Minimal Occurrences with Constraints and Time Lags). Separate antecedent and consequent inclusive constraints and window width are used by the MOWCATL approach to find sequential patterns and separated by the time lag. Algorithm returns the temporal rules with minimum confidence threshold values. Inge Jonassen et al. identified the conserved patterns form a set of unaligned protein sequences in [15] by proposing new methods. Those identified patterns are refined by guaranteed refinement algorithms to get only the flexible patterns and known motifs are recovered for PROSITE families. Assembling the frequent event which are close to each other in a database that are partially ordered is called as an episode predicts and describes the behaviour of the system. Two efficient WINEPI and MINEPI algorithms are proposed in [19] by Heikki Mannila finds all episodes and discovers only frequent episodes based on the minimal occurrences of the previous episode. In [21] Jian Pei et al. proposed a pattern-growth approach based on projection by converting the large sequence database into projected database and the frequent items are explored to develop the sequential patterns.

Mohammed J.Zaki reported that the SPADE algorithm works in the best way in [25] by splitting the larger problem into small sub-problems to solve them independently in the memory using the lattice search techniques and joined again using the join operations. In [24] Yanchang Zhao et al. presented impact oriented negative sequential rules to consider positive and negative correlation. SPAM (Sequential Pattern Mining) algorithm is used as a starting point. Two metrics such as contribution and impact have been considered to discover impact oriented rules. Innovative approach SOMAD (Service Oriented Mining for Antipattern Detection) proposed by Mathieu Nayrolles in [20] to detect SOA antipattern. Strong association is detected and filtered using detection metrics. Phillippe Fournier Viger et al. in [8] used a RuleGrowth algorithm find rules based on the pattern growth approach. First find rules between items in the database then the expandleft and expandright procedure is followed by scanning the database.

Srivatsan Laxman et al. in [18] studied the algorithms of temporal data mining and presented the results regarding the pattern discovery methods. Investors in the stock market wish to predict the behaviour of the customers by understanding the relationship between items purchased by the customers from the market. Y.L Hsieh et al. in [14] understand the relationship by using data mining knowledge and techniques. Arthur Pitman in [22] focused on mining sequential navigation patterns from web. Spurious sequences are eliminated using the closed criteria. Tjorben Bogon et al. presented a new approach in [4] by combining the assistance functionalities to analyse the input and output. EDASIM tool is developed for input selection, preparation and validation to assist the output data. Maureen A.Sartor et al. developed a web based gene set called ConceptGen and mapping tool in [23]. It offers more concepts related to the biological systems.

The previous research works offers limitations such as infrequent items leads to predict inaccurate data, rules have same support and confidence value rules generated are more and slow generation of rules. The proposed work overcomes

these issues by using the fuzzy logic technique that clearly removes all those infrequent items and provides fast prediction with accurate data.

III. PROPOSED WORK

A. Basic Terminologies

1) Accuracy: Measure of Predictive model that reflects the number of times that the model is correct when applied to the data.

2) Association rule: The rule in the form of if then else which associates the items in the database. Association is the relationship between two variables. For example, in the supermarket purchased items have association. An association rule is defined as the implication, X => Y.

3) Fuzzy logic: A logic based on the fuzzy set. It is a multivalued logic with representing the truth value on the closed interval [0, 1], where 0 is equated with the classical false value and 1 is equated with the classical true value.

4) Fuzzy set: A set of items with degree of membership in the set ranging from 0 to 1.

5) Fuzzy system: Set of rules that are formed using the linguistic variables described and processed by fuzzy sets and fuzzy logic.

6) Predictor: Used to build a predictive model to predict the data.

7) Predictive model: The model is created to perform the prediction.

8) Sequential pattern: Pattern analyzed from the items available in the sequence database. Based on this pattern the behavior, data can be predicted.

9) Sequential rule: Rules formed with the sequence database items. The rules are found with the frequent items.

10) Sliding window constraint: Window that moves from the beginning of the sequence to end of the sequence. The rule found with this constraint shows the time sequence.

B. Architecture

The proposed work predicts the accurate data from the huge amount data contained in the database. The process of prediction is classified as scanning the database, extracting the frequent items, forming rules for prediction with frequent items, finding rules occurring within the maximum amount of time, assigning weight value for each rule. This process can be briefly explained as rules are formed to predict the accurate value. With the strict ordering of rules partial matching does not help for making prediction. This can be addressed by partial ordering of the items where items in antecedent and consequent are unordered. Fuzzy technique is used to remove infrequent items from the database by classifying the data set using the linguistic terms and membership function represents the Boolean values. Frequent items alone taken to form rules using the RuleGrowth algorithm that scans the database to generate rule with the 1*1 size. Next the rule is expanded on both sides by following expandleft and expandright procedure by scanning the database further. The rule i=>j and j=>i are finally found with its support and confidence value. The rules found using this RuleGrowth algorithm are more. To find a rule that occurs within the maximum amount of time TRuleGrowth algorithm is proposed extension of the RuleGrowth algorithm accepting the sliding window constraint. Association rules are found by understanding the relationship between two variables. Weight values with certain range are assigned to the association rule which quickly predicts the accurate data needed by the user. Fig.1. is the prediction of accurate data.

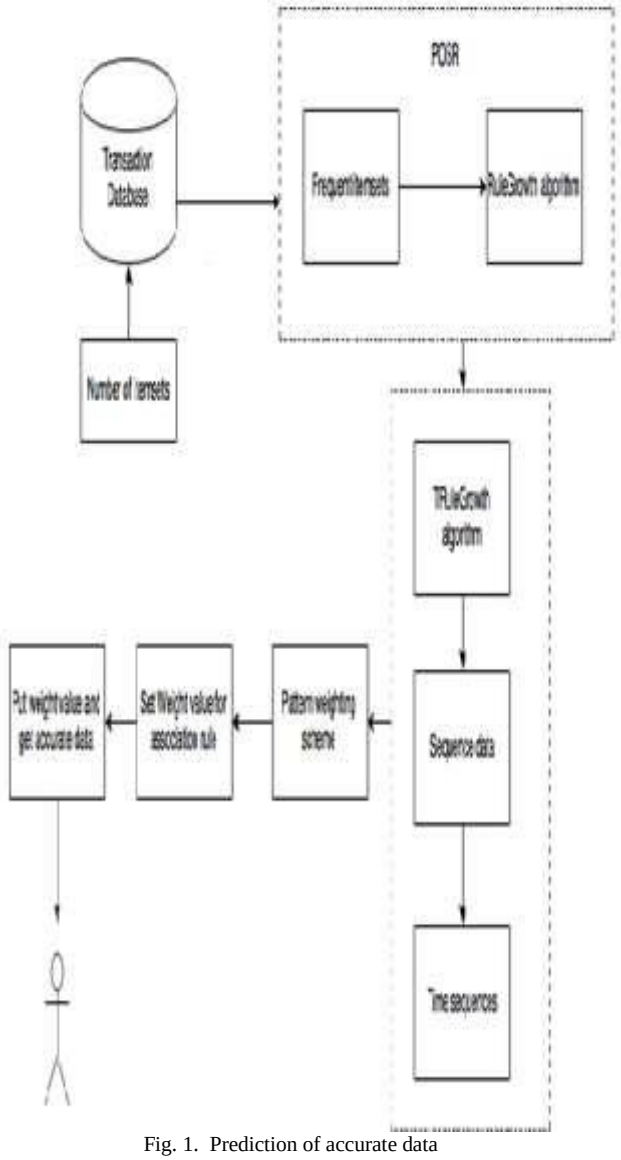

Fig. 1. Prediction of accurate data

System architecture is the conceptual model that describes the structure and behaviour of the system. Fig. 1. Represents the architecture of the proposed work. First, the transaction database is taken as the input. From this database the only frequent items are extracted by removing the infrequent items using the fuzzy logic technique. To eliminate infrequent items, linguistic terms which classify the database using and membership functions are used. The association rules are formed using the frequent items that have been extracted. The rules are found by following the procedure of ExpandRight and ExpandLeft the items in both antecedent and consequent. RuleGrowth algorithm possesses this expand procedure but more rules are formed which leads to confusion. Therefore, constraints are added to the algorithm and now there is an extension of RuleGrowth algorithm named TRuleGrowth algorithm find rules occurring with the sliding window constraint. Because users often wish to find rules that occur within the maximum amount of time. Finally weight value is assigned for each association rule and by using that weight value the accurate data is predicted.

C. Sequential Rule Extraction For Classification

A Sequence rule is applied to list all possible frequent items from the transaction database One database is given as the input to generate the sequential rule before classifying the dataset to form association rule. Fig.2. represents the sequential rule extractions where the administrator stores the data in the transaction database where the data are collected for the process for predicting the accurate data. After extracting data the id is given to the each transaction. Set the range value classify low, middle high value of attribute 1 classify low, middle high value of attribute2 remove infrequent item set linguistic terms set membership function classify the data into low high Apply sequential rule.

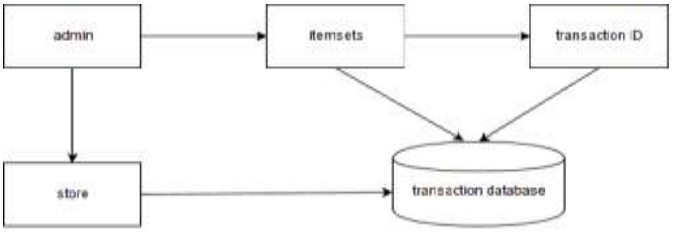

Fig. 2. Sequential Rule Extractions

D. Remove Infrequent Itemset

In this set the range value of each attributes for classifying the data set. Classify the data into low, middle, high values. Get the infrequent items and remove from the table. Fig. 3. Explain the process of removing infrequent items.

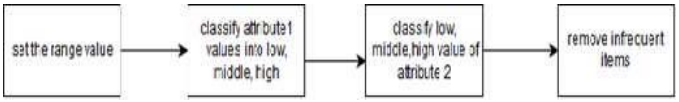

Fig. 3. Removing Infrequent Items

First the range value is given to the data set. Next based on that value the data set needs to be classified into low value, middle value and high value for both attribute that is attribute
1 and attribute 2. From this classification the infrequent can be removed and the frequent items only considered for the prediction.

E. Candidate Rule Prescreening

In this linguistic terms and membership function are set to classify the data set to apply it into the sequential rule which in turn used to predict accurate data. Membership functions are symbolic representation of each attribute. Fig.4. shows the candidate rule pre-screening.

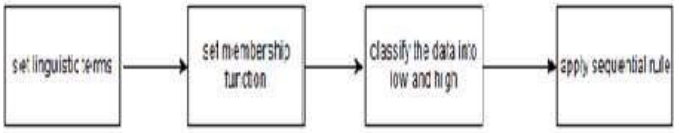

Fig. 4. Candidate Rule Pre-screening

The linguistic variable takes on linguistic values which are linguistic terms with the associated degrees of membership in the set. Transaction Database Frequent Itemsets sequential Rule mining Association Rules.

F. POSR Classification And Frequent Item

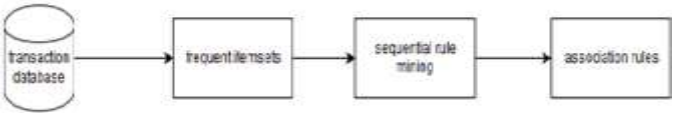

Fig. 5. POSR Classifications and Frequent Itemsets

In this get the frequent item set is obtained and by applying sequential rule mining algorithm association rules are generated. Fig.5. represent the partially ordered sequential rule classifications and the frequent item. With the infrequent item predicting the accurate data will be more difficult. Therefore, by using the fuzzy logic the infrequent item can be removed. This diagram shows that the transaction database is taken as the input which contains infrequent items as well as frequent items. Form this infrequent items get eliminated and the frequent item only considered for the mining the sequential rule and finally the association rules are formed for the prediction.

G. Algorithms

1) RuleGRowth algorithm

Algorithm: RuleGrowth

Input: database, minimum support, minimum confidence Output: sequential rule with their confidence and support Begin

Step 1: database is taken as the input

Step 2: algorithm scans the database once

Step 3: sequences are recorded with their sequence
*ID Step 4: find rule of size 1*1*
*Step 5: after finding the rule with size 1*1 the database is again scanned to find frequent rule by following the expanding procedure on both sides (EXPANDLEFT AND EXPANDRIGHT)*

Step 6: the expand procedure compared with the minimum confidence given as the input

Step 7: the expand procedure rule is calculated by dividing with the sequence database items

Step 8: if it is greater than minimum confidence the rule
{i}=>{j} and {j}=>{i} is return with their support and confidence value

End

The main procedure is to scan the database and generate rule of size 1*1. This rule is extended by scanning the database further to form large valid rules till the algorithm covers all the items in the database. EXPANDLEFT and EXPANDRIGHT procedure are applied to both sequences (i=>j and j=>i). The procedure value should be greater than are equal to minimum confidence to output the valid rule. This algorithm may be optimized by tracking the first and last occurrence in the sequence rule.

2) TRule Growth algorithm

TRule Growth algorithm is the extension of RuleGrowth that accepts the sliding window constraint to find rules occurring within the maximum amount of time. Two modifications are made to the procedure of RuleGrowth algorithm to form the TRuleGrowth algorithm. First, instead of considering the first and last occurrences of each item for each sequence, all occurrences of each item are considered for each sequence. Second, changing the procedure to check whether the i occurs before j and j occurs before i. sliding window constraint is applied from the starting rule itself size 1*1.

3) Fuzzy Logic

Fuzzy logic is the technique used to remove those infrequent items from the database with linguistic terms and membership function. Linguistic terms is the process of converting the variable into words and the membership function represents the data set using the Boolean value.

Weight value assigned to the association rule found to make the process of predicting easier. Best quality rules alone extracted finally by using those weight value reduction from larger rules.

Input: transaction database

Output: frequent items
Start

Step 1: database is taken as the input

Step 2: the database is classified based on the attribute range values

Step 3: after classification the variables are converted into words using the linguistic term as low, middle and high

Step 4: the membership function with the Boolean value '0' and '1' is used to again classify the data set as low, high and medium by converting the linguistic terms into Boolean variables.

Step 5: removes the infrequent from the data set

Step 6: the association rule forming procedure proceeds with frequent items are further it can be reduced by using the weight to form best quality rules.

Stop

IV. CONCLUSION

The two algorithms RuleGrowth and TRuleGrowth for the proposed work of partially ordered sequential rule find rules with excellent scalability and extension of the RuleGrowth algorithm is the TRuleGrowth which accepts the sliding window constraint. Eliminating the infrequent items before generating the sequential rules so that rules are found only with the frequent items. Using fuzzy logic infrequent items are removed which improves the performance and also the prediction accuracy. Weight value assigned to the association rules also improves the accuracy of the predicting data. Best quality rules only extracted finally which increases the speed of the process. In future it can be extended to predict particular value from the set of predicted data. For that obtained data threshold value can be given for the weight value assigned for the association rule as minimum support threshold value and maximum confidence threshold value. Y assigning threshold value the very particular data can be obtained.

REFERENCES

[1] R. Agrawal, T. Imielminski, and A. Swami, ―Mining association rules between sets of items in large databases,‖ in Proc. 13th ACM SIGMOD Int. Conf. Manage. Data, 1993, pp. 207–216.
[2] R. Agrawal and R. Srikant, ―Mining sequential patterns,‖ in Proc. 11th International Conference of Data Engineering., 1995, pp. 3–14.
[3] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, ―Sequential Pattern mining using a bitmap representation,‖ in Proc. 8th ACM

Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 429–435.
[4] T. Bogon, I. J. Timm, A. D. Lattner, D. Paraskevopoulos, U. Jessen, M. Schmitz, S. Wenzel, and S. Spieckermann, ―Towards assisted input and output data analysis in manufacturing simulation:The EDASIM approach,‖ in Proceedings of Winter Simulation Conference, 2012, pp. 257–269.
[5] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth, ―Rule discovery from time series,‖ in Proc. 4th ACM Int. Conf. Knowl. Discovery Data Mining, 1998, pp. 16–22.
[6] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. Mephu Nguifo, ―CMRules: An efficient algorithm for mining sequential rules common to several sequences,‖ Knowledge Based System, Elsevier, vol. 25, no. 1, pp. 63–76, 2012.
[7] P. Fournier-Viger and V. S. Tseng, ―TNS: Mining Top-K Nonredundant sequential rules,‖ in Proc. 28th Symp. Appl. Comput., ACM, 2013, pp. 164–166.
[8] P. Fournier-Viger, R. Nkambou, and V. S. Tseng, ―RuleGrowth: Mining sequential rules common to several sequences by patterngrowth,‖ in Proc. 26th ACM Symp. Appl. Comp., 2011, pp. 954–959.
[9] P. Fournier-Viger, T. Gueniche, and V. S. Tseng, ―Using partiallyordered sequential rules to generate more accurate sequence prediction,‖ in Proc. 8th Int. Conf. Adv. Data Mining Appl., Springer, 2012, pp. 431–442.
[10] P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, ―Fast vertical sequential pattern mining using co-occurrence information,‖ in Proc. 18th Pacific-Asia Conf. Knowl. Discovery Data Mining, Springer, 2014, pp. 40–52.
[11] P. Fournier-Viger, A. Gomariz, A. Soltani, T. Gueniche, C. W. Wu, and V. S. Tseng, ―SPMF: A Java open-source pattern mining library,‖ Journal of Machine Learning Research., vol. 15, pp. 3389–3393, 2014.
[12] S. K. Harms, J. Deogun, and T. Tadesse, ―Discovering sequential association rules with constraints and time lags in multiple sequences,‖ in Proc. 13th Int. Symp. Method. Intell. Syst., Springer, 2002, pp. 373–376.
[13] H. Minqing and B. Liu, ―Opinion feature extraction using class sequential rules,‖ presented at the AAAI Spring Symp. Computational Approaches Analysing Weblogs, Palo Alto, USA, Mar. 2006.
[14] Y. L. Hsieh, D.-L. Yang, and J. Wu, ―Using data mining to study upstream and downstream causal relationship in stock market,‖ in Proc. 9th Joint Conf. Inf. Sc., ACM, 2006.
[15] I.Jonassen, J. F. Collins, and D. G. Higgin, ―Finding flexible patterns in unaligned protein sequences,‖ Protein Sci., Journal of Computational Biology, vol. 4, no. 8, pp. 1587–1595, 1995.
[16] D. Lo, S.-.C. Khoo, and L. Wong, ―Non-redundant sequential rules—Theory and algorithm,‖ Inf. Syst., Esevier, vol. 34, no. 4/5, pp. 438– 453, 2009.
[17] D. Lo, G. Ramalingam, V. P. Ranganath, and K. Vaswani, ―Mining quantified temporal rules: Formalism, algorithms, and evaluation,‖ in Proc. 16th Working Conf. Reverse Eng., 2009, pp. 62–71.
[18] S. Laxman and P. Sastry, ―A survey of temporal data mining,‖ Sadhana, vol. 3, pp. 173–198, 2006.
[19] H. Mannila, H. Toivonen, and A. I. Verkano, ―Discovery of frequent episodes in event sequences,‖ Data Mining Knowl. Discovery, vol. 1, no. 3, pp. 259–289, 1999.
[20] M. Nayrolles, N. Moha, and P. Valtchev, ―Improving SOA antipatterns detection in service based systems by mining execution traces,‖ in Proc. 20th IEEE Working Conf. Reverse Eng., 2013, pp. 321–330.
[21] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, ―Mining sequential patterns by pattern-growth: The prefixspanapproach,‖ IEEE Transaction of Knowledge and Data Engineering, vol. 16, no. 10, pp. 1–17, Oct. 2004.
[22] Pitman and M. Zanker, ―An empirical study of extracting multidimensional sequential rules for personalization and recommendation in online commerce,‖ in Proc. 10th International Conference on Wirtschaftinformatik, 2011, pp. 180–189.
[23] M. A. Sartor, V. Mahavisno, V. G. Keshamouni, J. Cavalcoli, Z. Wright, A. Karnovsky, R. Kuick, H. V. Jagadish, B. Mirel, T. Weymouth, B. Athey, and G. S. Omenn, ―ConceptGen: A gene set enrichment and gene set relation mapping tool,‖ Bioinformatics, vol. 26, no. 4, pp. 456–463, 2010.
[24] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, ―Mining both positive and negative impact-oriented sequential rules from transactional data,‖ in Proc. 13th Pacific-Asia Conf. Knowl. Discovery Data Mining, Springer, 2009, pp. 656– 663.
[25] M. J. Zaki, ―SPADE: An efficient algorithm for mining frequent sequences,‖ Mach. Learning, ACM, vol. 42, no. 1–2, pp. 31–60, 2001.