IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-4, Issue-11 E-ISSN: 2347-2693

Use of Constraints in Pattern Mining: A Survey

R.V. Mane[1*], V. R. Ghorpade[2]

[1*]Department of Technology, Shivaji University, Kolhapur, Maharashtra, India
[2]D.Y.Patil College of Engineering and Technology, Kolhapur, Maharashtra, India

E-mail: rvm_tech@unishivaji.ac.in, vijayghorpade@hotmail.com

Available online at: www.ijcseonline.org

Received: Oct/19/2016 Revised: Oct/30/2016 Accepted: Nov/19/2016 Published: Nov/30/2016

***Abstract*—Constraint based pattern mining and association rules are used in many applications like genetic sequence analysis, in finance for bankrupting prediction, in securities for fraud detection, in agriculture for discovering classification of plants etc. to get the user interesting knowledge. Constraints are useful to eliminate unwanted rules and also solves rule explosion problem. Many algorithms are proposed for constraint based pattern mining and association rule generation. These constraints are in the form of attribute, item length, time or duration, regular expression etc. Pushing constraints in a mining process gives user interesting discovery. Literature survey shows that performance of an algorithm improves with application of constraint during the mining process. The paper elaborates about the literature survey on use of constraints in generation of association rules with different categories of constraints with its properties.**

Keywords-constraint;frequent;sequence;pattern;mining

I. INTRODUCTION

Pattern mining is used for finding user interesting knowledge, for prediction or for classification. Numbers of applications are used and various algorithms are designed for finding associations or co-relations from patterns. But the major drawback of pattern mining algorithms is either with performance or with its functionality. Constraint based mining is used to develop systematic approach to sequential pattern mining. With adding constraints at the source level or during post-processing gives appropriate, strong and valid, user interesting information. This information in the form of discovered rules or patterns are appropriately used for getting the knowledge or for prediction purpose [15].

Constraint based mining is used in every approach of finding frequent itemsets. With use of Apriori principle sequence pattern mining algorithms as GSP[1], SPAM [2], SPADE[3] are used with time constraints like *minGap*, *maxGap*, window size etc. These Apriori algorithms are also used with item constraint as attributes, item length, regular expression etc. Attribute constraints can be used with Pattern growth approaches such as PrefixSpan [4], FreeSpan [5]. Main aim of constraint based pattern mining is to get user interesting patterns .These constraints are in the form of different values of itemsets, gap among the data items, time, length or aggregate values.

Paper does the survey of constraints based pattern mining algorithms. First part of the paper gives the details of constraints in the form of its type and categories based on their properties and second part does the survey of algorithms with different constraints as Regular Expression constraint, attribute constraint and time or gap constraint during discovery of knowledge in the form of patterns. Literature review in the paper highlights an idea as how efficiency of an algorithm will be increased with pushing constraints in the mining process and useful for finding user interesting information.

[*]Corresponding Author:
Rashmi V.Mane
E-mail: rvm_tech@unishivaji.ac.in

II. TYPES AND CATEGORIES OF CONSTRAINTS

The strategy which allows users to specify their expectations in the form of constraints to confine the search space is called constraint based mining. These constraints are of different types [6] as

Knowledge type-It specifies the type of knowledge to be mined in the form of association, classification or clustering.
Data- It gives the set of data items.
Dimension or level-It gives the specific attributes or its abstraction level.
Interestingness-It is in the form of statistical measures such as support, confidence or correlation etc.
Rule constraint-It is expressed in the form of maximum or minimum number of predicates, antecedents, consequences or relation among attributes.

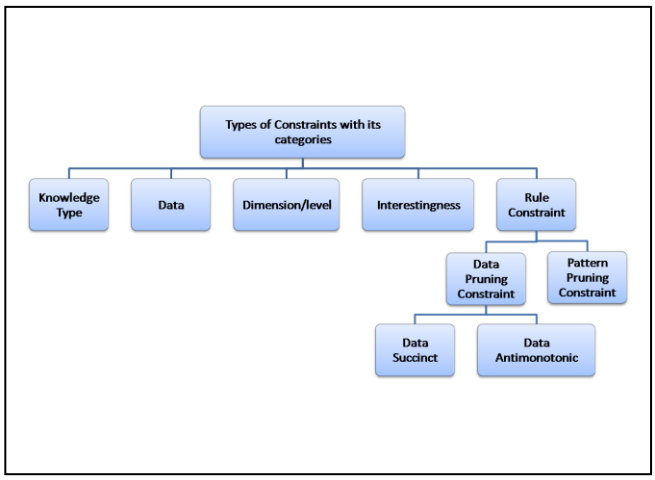

Figure 1. Types of Constraints with its categories

These data constraints are categorized into different types of constraints are shown as follows-

***Constraint 1: Item Constraint* –** It specifies a subset of items that should or not be present in pattern.

Constraint 2: Length constraint- It specifies the requirement of the length of pattern.

Constraint 3: Super pattern Constraint- It is in the form of set of patterns .To find patterns that contain a particular set of pattern as sub pattern.

Constraint 4: Aggregate Constraint-It is on aggregation of item in a pattern which aggregation function can be Sum, Avg, max, min, standard deviation etc.

Constraint 5: Regular Expression Constraint-It is the constraint specified as a regular expression over set of items mining regular expression operators such as Kleene's closure etc.

Constraint 6: Duration Constraint-It is defined only in sequential database where each transaction in every sequence has a time stamp. It requires that time stamp difference between first and last transaction in a sequential pattern must be shorter or larger for a given period.

Constraint 7: Gap Constraint-It is defined only in sequence database where each transaction in every sequence has timestamp.

***Time Gap Constraint* –**It gives time interval between two adjacent elements to a reasonable period.

Categorization of Constraints:

Based on the properties of constraints Pattern pruning constraints are categorize into types as Antimonotone, Monotone, Succinct ,Convertible which is further classified as convertible antimonotone , convertible monotone , strongly convertible and Inconvertible constraints [6].

Constraints are categorized based on montonicity, antimonotonicity, succinctness.

Antimonotonicity- A constraint is antimonotonic for a sequence α implies that every non empty subsequence of α also satisfies this constraint.

This property allows Apriori algorithm to prune significant number of candidate sequence which require support counting. These constraints are restricted to iterative pruning .Confidence does not have antimonotone property but confidence of rules generated from same itemset has an antimonotone property.

Monotonic- A constraint is monotonic for a sequence α implies that every super sequence of α also satisfy this constraint.

Succinct-It is specified with precise formula. Item and item length constraint are having antimonotonicity and succinct property. Super pattern and aggregate are satisfying monotonic and succinct property.

Prefix Anti-Monotonic- A constraint is prefix-anti-monotonic for each sequences α, so does every prefix of α.

Prefix Monotonic-A constraint is prefix-monotonic for each sequences α so does every sequence having α as a prefix.

A constraint is called prefix-monotonic if it is prefix-antimonotonic or prefix monotonic. If length constraint is antimonotonic then it must be prefix antimonotonic.

III. USE OF CONSTRAINTS IN SEQUENTIAL PATTERN MINING

A. Regular Expression Constraint

Importance of the constraints is essential for many sequential patterns mining applications proposed **by Jian Pei, Jiawei Han, and Wei Wang** in "**Constraint based sequential pattern mining the pattern growth methods**". In this paper authors have conducted a systematic study on constraint based sequential pattern mining. The main focus is given to user interesting sequential patterns. Firstly they have given a classification of constraint based on their application and their role in Sequential Pattern Mining. Secondly they have specified a new framework for prefix-monotone property for the constraints like regular expression.

Paper gives the idea of mining sequence pattern with Regular expression. During recursive mining if a prefix itself is a pattern which satisfies the user interesting constraint then it should be outputted. Further this prefix satisfying constraint should be grown and mined recursively. The process ends when there is no local frequent item or no legal prefix.

Experimental study of paper shows that although prefix monotone property is weaker than Apriori but it still achieves better performance than Apriori based method. The paper

explores the pushing method of aggregate constraint in pattern growth approach. An item is considered as a small item if its value is ≤ v, where v is the user specified threshold. Otherwise it is called a big item. In the first scan of projected database, unpromising big items are removed. In α projected database, when a pattern β is found following α as small item, first it is checked whether small item can be replaced by a big item x and still found average value satisfying constraint. Then prefix <α, x> is marked as promising and not checked again. If <α, x> violates the constraint then projected db can be pruned. This is called unpromising pattern pruning rule. Author states that pattern growth can be used to handle some tough aggregate constraint without prefix-monotone property [7].

"**SPIRIT: Sequential Pattern Mining with Regular Expression Constraint**" proposed by **Minos Garofalakis, Rajeev Rastogi and Kyuscok Shim**. They have elaborated sequential pattern mining algorithm's unfocused approach. Two major drawbacks of pattern mining algorithms are given as firstly disproportionate computation cost for selective users and overwhelming volume of potentially useless results. To overcome this problem, they have proposed an algorithm which incorporates user controlled focus in mining process.

Author have presented the problem of mining sequential patterns with Regular expression constraint and pushed this constraint inside pattern mining process. Algorithm exploits equivalence of regular expression to deterministic finite state automata. The experiment study shows that including regular expression into pattern mining computation yield more improvement in performance. Four SPIRIT algorithms points spanning the entire spectrum of relaxation for user specified Regular expression as follows-

SPIRIT (N)-N for Naive. It employs weakest relaxation of regular expression. It prunes only candidate sequence containing elements that don't appear in Regular expression.

SPIRIT (L)-L for Legal. It requires every candidate sequence to be legal with respect to some state of automata.

SPIRIT (V)-V for valid. It filters out candidate sequence that is not valid with respect to any state of automata.

SPIRIT(R)-R for regular. It pushes regular expression inside mining process by counting support only for valid candidates [8].

Leticia Gomez, Alejandro Vaisman has proposed a language based on regular expression to restrict frequent sequence to user specified constraints. The language they referred as RE-SPaM based on constraints over atomic item. Expression in the language contains attributes, functions over attribute and variables. They have used moving objectdb (MOD) which include trajectory aggregation in traffic analysis. Example provided in the paper as tourist application in the city of Paris where POI may be restaurant, hotel, tourist attraction etc. In Re-SPaM constraints are enclosed in square brackets. These constraints will also contain function over attributes such as rollup [14].

B. Attribute Constraint

Shigeaki Sakurai, Youichi Kitahata and Ryahei Orihama proposed a method on **"Discovery of Sequential patterns based on constraint patterns"**. This method proposes user interest as constraint patterns. The paper deals with dependent items where attributes show group of items that have a common property. Paper introduces constraint patterns in order to extract valid sequence patterns. These patterns are further used to predict sub patterns of it. These constraint patterns are reformed as constraint item sequence. If length of constraint pattern is large, combination of attribute values also increases exponentially so proposed method checks these constraint itemset step by step.

The method first generates candidate itemset and finds whether it is frequent or not then repeats the process for candidate sequence patterns .Process repeated till all frequent sequential patterns are discovered. A candidate item set with (i+1) items are generated from two frequent items with i item. These frequent itemsets have to satisfy the constraint item subset. In the next step $(k+1)^{th}$ candidate sequence pattern is generated from two k^{th} frequent sequential patterns. It must satisfy constraint pattern. This decomposition of constraint patterns is done with user given some constraint patterns. These patterns are decomposed into constraint sub patterns and constraint item subsets respectively. These decomposed constraint patterns are used for generation step of sequential pattern mining method. If candidate doesn't satisfy constraint, it will not calculate its support. The authors have verified effectiveness of this method on the sequence data of stock price indexes [9].

C. Time and Gap Constraint

Constraint based mining of sequential patterns is an active research area given by **Mario Leleu,Christophe Rigotti, Jean-Francois Boulicaut and Guillaume Euvrend.** In the paper "**Constraint based mining of sequential patterns over dataset with consecutive repetitions**", Author has proposed an algorithm which computes consecutive repetition in sequence dataset by reducing the amount of data to process which speeds up the extraction time. They have added a time window constraint and constraint of sequence length. To reduce the size of occurrence list they have used the strength of a constrained generalized occurrence list, which contain identifiers, timestamp of an occurrence of first event, interval of min-max and timestamp of last occurrence of last event of pattern. They have proposed GoSpec algorithm which is an instance of abstract algorithm with joint designed for generalized occurrence list. Author has carried out the experimentations with cSpade algorithm.

Experiments show that with GoSpec algorithm, the gain in terms of memory space and execution time is achieved [10].

Yen-Liang Chen, Ya-Han Hu has proposed a new constraint in the form of recency in the paper **"Constraint based sequential pattern mining: The consideration of recency and compactness."**Recency causes patterns to adapt to latest behaviours in sequential database and compactness gives reasonable time spans for these patterns. The author refers these patterns as CFR-patterns.

Paper proposes CFR-postfixSpan algorithm which is developed from PrefixSpan algorithm for finding frequent sequence patterns with additional two constraints as recency and *ms-length*. Recency minSup for recent sequence database gives the most recently occurring subset of sequential database and compact constraint means the time span from first item to last item in a pattern must be no more than maximum span length or *ms_length*.

Author has done comparison in PrefixSpan and CFR-postfixSpan algorithm. First difference is PrefixSpan is concerned with order of items in data sequence while CFR handles timestamp of each item in data sequence. Secondly CFR projects only frequent recent postfixes.

Algorithm starts with finding all length-1 CFR patterns. For each item *cf_support* and *cr_support* value is calculated. In the second step it divides the search space. For each compact projected database respective *sid* and *Endtime* values are recorded and compactness constraint is satisfied. Compactness constraint is the difference between timestamp of earliest item and *Endtime* <= *ms_length*. In the final step all CFR-patterns can be found by constructing corresponding projected database and recursive mining each one.

Author's experimental study shows that by adding *r-minsup* and *ms-length* properly, many uninteresting patterns can be pruned and CFR-postfixSpan performs well for long sequence database [11].

Tedeusz Morzy, Marek Wajeiechowski, Maciej Zakrzewicz has discussed dataset filtering techniques for "**Efficient Constraint based Sequential Pattern mining using dataset filtering techniques**". Author has given an extension of GSP algorithm for dataset filtering. They have proposed GSP-F algorithm by incorporating filters in GSP algorithm. As GSP algorithm generate candidate sequence iteratively and computes support or occurrence of data sequence form its source dataset. In GSP-F filtering is done in iteration. Those patterns not satisfying constraint are not included in candidate verification process. In the post processing step all those patterns are filtered which don't satisfy user specified pattern constraint. Experimental study of the given research work states that lower the selectivity of dataset filtering constraint, better the performance of GSP-F than GSP [12].

Yu Hirate, H.Yamana has proposed generalize sequential pattern mining with item interval in a paper "**Generalized sequential pattern mining with item interval**". The algorithm includes three points as a capability to handle item gap and time interval, a capability to handle extended sequences and adopting four item interval constraints. According to author this proposed method is able to substitute all types of conventional sequential pattern mining algorithms with item intervals. Using Japanese earthquake data, they have confirmed that algorithm is capable to discover sequential patterns with item interval, defined in a flexible manner by the interval itemization function [13].

IV. CONCLUSION

Literature survey shows that many algorithms are proposed for constraint based pattern mining .Constraints is pushed in pattern mining algorithms. Accuracy in the result and solving real life problems can be achieved with adding constraints in the mining process. Survey shows that adding constraints in the mining process improves efficiency of an algorithm and it decreases execution time by pruning the search space. Result produced by algorithms are more easy to analyze and small in number which increases prediction accuracy. Many real life problems like detection of heart disease, pattern structure analysis etc. are solved with pushing constraints in pattern mining.

REFERENCES

[1] Ramakrishnan Srikant and Rakesh Agrawal,"Mining Sequential Patterns: Generalizations and Performance Improvements‖", Advances in Database Technology — EDBT '96 Lecture Notes in Computer Science, Springer, Volume 1057, pp 1-17, 1996

[2] Ayres, J.,Gehrke, J.Yiu T. and Flannick J.' Sequential Pattern Mining using Bitmap representation in proceeding of ACM SIGKDD'02,pp 429-35,2002

[3] M.J.Zaki, "Spade: An efficient algorithm for mining frequent sequences", Machine Learning 42 (2001), pp. 31-60

[4] J.Pei, J han, B.Mortazavi PrefixSpan: Mining Sequential Patterns efficiently by prefix projected pattern growth ,ICDE 2001,Heidelberg,Germany,2001 pg 215-224

[5] H.Han, J.Pei, B.Mortazavi- Asl,Q.Chen, U. Dayal and M-C Hsu(2000) FreeSpan: Frequent Pattern projected Sequential Pattern mining: Proceedings of 2000 International Conference on Knowledge discovery and data mining pp 355-359

[6] Jiawei Han,Micheline Kamber,Jain Pei, "Data Mining Concepts and Techniques",Third Edition,2012,ISBN:978-0-12-381479-1.

[7] J.Pei, J.han and W.Wang, "Constraint based sequential pattern mining: the pattern growth method", in journal of Intelligent Information systems (2007), pp.133-160.

[8] M.N.Garofalakis, R.Rastogi and K.Shim, " Spirit: Sequential pattern mining with regular expression constraints", In

Atkinson, M.P., Orlowska, M.E., Valduriez , P., Zdonik, S.B., Brodie, M.L., eds.: VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK, Morgan Kaufmann (1999),pp. 223-234 .

[9] Shigeaki Sakurai, Youichi Kitahara and Ryohei Orihana, "Sequential Pattern Mining based on new criteria and attribute constraint",IEEE International Conference on Systems,Man and Cybernetics 2007.

[10] Marion Leleu,Christophe Rigotti,Jean-Francois Boulicar and guillaume Euvrard, "Constraint Based Mining of Sequential Patterns over Dataset with Consecutive Repetitions",PKDD 2003.

[11] Yen-Liang Chen,Ya-Han Hu, "Constraint Based Sequential Pattern Mining:The Consideration of Recency and Compactness ",Decision Support System by Elseveir 42(2006) pp.1203-1215.

[12] Tedeusz Morzy,Marek Wajeiechowski,Maciej Zakrzewicz," Efficient Constraint based Sequential Pattern Mining using dataset filtering techniques", Proceedings of the Baltic Conference, BalticDB&IS 2002, Volume 1,2002

[13] Yu Hirate, H.Yamana, "Generalized sequential pattern mining with item interval",Proceedings of 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2006.

[14] L.Gomez, A.A.Vaisman, "Re-SPAM: Using Regular Expression for sequential pattern mining in Trajectory database", IEEE International Conference on Data Mining workshops 2008

[15] D.Senthil Kumar , N.Jayaveeran, " A survey on Association Rule Mining Algorithms for Frequent Itemset ",International Journal of Computer Science and Engineering,Vol.4,Issue-10,Oct-2016

Authors Profile

R V Mane pursed Bachelor of Engineering from Shivaji University,Kolhapur,India and Master of Technology from Shivaji University,Kolhapur,India in year 2009. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Technology, Department of Computer Science and Technology, Shivaji University,Kolhapur, Maharashtra,India. She has published more than 10 research papers in reputed international journals and conferences including IEEE,Elsvier,Springer and it's also available online. Her main research work focuses on Data Mining.

V R Ghorpade currently working as Principal in D.Y.Patil College of Engineering and Technology,Kolhapur,Maharashtra.He has completed Ph.D. in Computer Science and Engineering.He is a member of IEEE & IEEE computer society. He has published more than 20 research papers in reputed international journals and conferences including IEEE and it's also available online. His main research work focuses on Network Security, Data Mining,Adhoc Networks.

