ISSN:2320-0790

An International Journal of Advanced Computer Technology

FSRM: A Fast Algorithm for Sequential Rule Mining

Anjali Paliwal, Mr. Sourbha Dave
M.E. Student, Asst. Professor
, Medicaps Institute of Technology & Management Indore (M.P.)

ABSTRACT: Recent developments in computing and automation technologies have resulted in computerizing business and scientific applications in various areas. Turing the massive amounts of accumulated information into knowledge is attracting researchers in numerous domains as well as databases, machine learning, statistics, and so on. From the views of information researchers, the stress is on discovering meaningful patterns hidden in the massive data sets. Hence, a central issue for knowledge discovery in databases, additionally the main focus of this paper, is to develop economical and scalable mining algorithms as integrated tools for management systems.

I. INTRODUCTION

Data mining, that is additionally cited as knowledge discovery in databases, has been recognized because the method of extracting non-trivial, implicit, antecedently unknown, and probably helpful data from knowledge in databases. The information employed in the mining method usually contains massive amounts of knowledge collected by computerized applications. As an example, bar-code readers in retail stores, digital sensors in scientific experiments, and alternative automation tools in engineering typically generate remendous knowledge into databases in no time.

As computers are employed in a lot of areas, massive volumes of knowledge are collected and hold on within the database continuously. This sort of data includes the group action records in supermarkets, banks, stock markets, and telephone firms. With the increasing volume of the hold on information, a vital issue is to work out a way to notice the helpful info from this huge history information. Data mining, conjointly called knowledge discovery in databases, is such a hunt space to extract implicit, graspable, antecedently unknown and doubtless helpful information from data.

The explosive growth in hold on knowledge has generated an imperative need for brand spanking new techniques and automatic tools that may show intelligence assist us in remodeling the immense quantity of into helpful information and knowledge.

The discipline involved with this task is currently referred to as data mining. If we try and capture this idea into a proper definition, then we can able to} outline data mining as" the analysis of (often large) empiric data sets to search out unexpected relationships and to summarize the information in novel ways in which are each comprehendible and helpful to the data owner".

II. BACKGROUND & RELATED WORK

ASSOCIATION RULE MINING

Association rule mining [1] is a popular knowledge discovery technique for discovering associations between items from dealing information. Formally, a transaction information D is defined as a set of transactions T= and a collection of items I=, where $t1,t2,\ldots,tn \subseteq I$. The support of an itemset X in I for a database is denoted as sup(X) and is calculated because the range of transactions that contains X. the matter of mining association rules from a transaction information is to seek out all association rules $X \rightarrow Y$, such that X,Y belongs I, $X \cap Y$=null, and that the rules respect some nominal marginal criteria. the two powerfulness criteria at first proposed by Agrawal [1] are that well-mined rules have a support larger or equal to a user-defined threshold minsup and a confidence greater or equal to a user-defined threshold minconf. The support of a rule $X \rightarrow Y$ is outlined as sup(X Y)/ |T|. the confidence of a rule is defined as $conf(X \rightarrow Y) = sup(X\ Y) / sup(X)$.Since $|T| \geq sup(X)$ for any X in I, the relation $conf(r) \geq sup(r)$ hold for any association rule r.

SEQUENTIAL RULE MINING

Association rules are mined from transaction databases. A generalization of transaction information that contains time info regarding the occurrence of items could be a sequence database (by Agrawal & Srikant [2]). A sequence database SD is outlined as a group of sequences S= and a group of items I=, wherever every sequence sx is an ordered list of transactions sx={X1, X2, … Xn}specified X1, X2, …Xn in I .

We propose the subsequent definition of a sequential rule to be discovered in sequence information. A sequential rule X =>Y is a relationship between two itemsets X,Y specified X,Y in I and X∩Y = null. The interpretation of a rule X=>Y is that if the items of X occur in some transactions of a sequence, the items in Y can occur in some transactions afterwards from identical sequence. Note that there's no ordering restriction between items in X and between items in Y. we tend to define two powerfulness measures for such a rule that are an adaptation for multiple sequences of the measures used for different sequential rule mining algorithms [4]. The primary measure is that the rule's sequential support and is defined as: seqSup(X => Y) = sup(XY) / |S|. The second measure is that the rule's sequential confidence and is defined as: seqConf(X => Y) = sup(XY) / sup(X). Here, the notation sup(XY) denotes the quantity of sequences from a sequence database wherever all the items of X seem before all the items of Y (note that items from X or from Y don't have to be within the same transaction). The notation sup(X) represents the number of sequences that contains X. Since |S| ≥ sup(X) for any X in I , the relation seqConf(r) ≥ seqSup(r) holds for any sequential rule r.

III. RELATED METHODS

Sequential rule mining has been applied in many domains like stock exchange analysis (Das & Lin [4], Hsieh, Wu & Yang [6]), weather observation (Hamilton & Karimi, [8]) and drought management (Harms & Tadesse, [7], Deogun & Jiang, [5]).

The most known approach for sequential rule mining is that of Mannila & Verkano [3] and alternative researchers later on that aim at discovering partly ordered sets of events showing often within a time window during a sequence of events. Given these "frequent episodes", a trivial algorithmic rule will derive sequential rules respecting a lowest confidence and support (Mannila [3]). These rules are of the shape X=>Y, where X and Y are 2 sets of events, and are taken as "if event(s) X seems, event(s) Y can possibly occur with a given confidence.

However, their work can only get rules during a single sequence of events. Alternative works that extract sequential rules from one sequence of events are the algorithms of Hamilton & Karimi [8], Hsieh, Wu & Yang [6] and Deogun & Jiang [5], that respectively discover rules between many events and one event, between 2 events, and between many events.

Contrarily to those works that discover rules during a single sequence of events, some works are designed for mining sequential rules in many sequences (Das & Lin [4]; Harms & Tadesse [7]). as an example, Das & Lin [4] discovers rules where the left a part of a rule can have multiple events, however the correct half still needs to contain one event. This may be a significant limitation, as in real-life applications, sequential relationships may involve many events. Moreover, the algorithmic rule of Das & Lin [4] is extremely inefficient.

IV. PRPOSED METHOD

INPUT

1: A source database D.

2: MST (Minimum Support Threshold).

3: MCT (Minimum Confidence Threshold).

PROCEDURE:

(1) We set the minimum support and scan the database to get 1-itemsets. In this step, we also count each item's support by using compressed data structure, i.e. head and body of the database. Here body of the database contain itemset with their support and arranges in the lexicographic order, i.e. sorted order. The proposed method first scans the sequential data base once & it count the support single size frequent items. the algorithm then arrange each pair of frequent elements in form of a rule. Like for a pair {1,2}, it generates two rules as : 1 => 2 & 2=>1. It also eliminates all infrequent items.

(2) Then, the sequential support and sequential confidence of every rule is calculated. All rules having support and confidence greater than the minimum threshold are valid rules.

(3) Then all the rules found in step 2 are expanded on their left and right side as follows:

Left-side growth is the method of taking two rules X=>Y and Z=>Y, wherever X and Z are itemsets of size n sharing n-1 items, generate a larger rule

X U Z => Y. Right-side expansion is the process of taking two rules Y=>X and Y=>Z, wherever X and Z are unit itemets of size n sharing n-1 items, to generate a larger rule Y U X=>Z. These two processes are applied recursively to seek out all rules ranging from rule of size 1*1 (for example, rules of size 1*1 permits finding rules of size 1*2 and rule of size 2*1).

V. RESULT ANALYSIS

Consider the following database.

Table 1: A Sequence Database [3]

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 result comparison of previous algorithm i.e. Previous Algorithm and proposed algorithm is shown in table 2 and will be as under:

Table 2: Result Comparison

Algorithm	Time	Space	No of Rules
CMRules	47 ms	0.89 Mb	9
Proposed	15 ms	0.55 Mb	9

VI. CONCLUSION:

In this paper, we tend to confer a completely unique algorithm for mining sequential rules. Unlike revious algorithms, it doesn't use a generate-candidate-and-test approach. Instead, it uses a pattern-growth approach for discovering valid rules specified it will be rather more efficient and scalable. It initial finds rules between 2 items and so recursively grows them by scanning the database for single items that would expand their left or right components.

VII. REFERENCES

[1] Rakesh Agrawal, Swami, A., & T. Imielminski, 1993, Mining Association Rules Between Sets of Items in Large Databases, *SIGMOD Conference*, pp. 207-216

[2] Rakesh Agrawal, & Ramakrishnan Srikant, 1995, Mining Sequential Patterns. Proc. Int. Conf. on Data Engineering, pp. 3-14.

[3] Mannila, H., Toivonen & H., Verkano, A.I. 1997. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(1): 259-289.

[4] King Ip Lin., Heikki Mannila, Gautam Das, Gopal Renganathan, & Padhraic Smyth, 1998. Rule Discovery from Time Series. *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining*.

[5] Liying Jiang & Jiternder S Deogun, 2005. Prediction Mining – An Approach to Mining Association Rules for Prediction. *Proceeding of RSFDGrC 2005 Conference*, pp.98-108.

[6] Hsieh, Y. L., Yang, D.-L. & Wu, J. 2006. Using Data Mining to Study Upstream and Downstream Causal Realtionship in Stock Market. *Proc. 2006 Joint Conference on Information Sciences*.

[7] Harms, S. K., Deogun, J. & Tadesse, T. 2002. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. Proc. 13th Int. Symp. on Methodologies for Intelligent Systems, pp. .373-376.

[8] Hamilton, H. J. & Karimi, K. 2005. The TIMERS II Algorithm for the Discovery of Causality. *Proc. 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 744-750.