Computers & Operations Research 52 (2014) 241–250

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Optimizing frequent time-window selection for association rules mining in a temporal database using a variable neighbourhood search

Yiyong Xiao [a], Yun Tian [a], Qiuhong Zhao [b,*]

[a] School of Reliability and System Engineering, Beihang University, Beijing 100191, China
[b] School of Economics and Management, Beihang University, Beijing 100191, China

ARTICLE INFO

Available online 5 October 2013

Keywords:
Data mining
Association rule with time-window
Integer programming
Variable neighbourhood search

ABSTRACT

In this study, we investigate the problem of maximum frequent time-window selection (MFTWS) that appears in the process of discovering association rules time-windows (ARTW). We formulate the problem as a mathematical model using integer programming that is a typical combination problem with a solution space exponentially related to the problem size. A variable neighbourhood search (VNS) algorithm is developed to solve the problem with near-optimal solutions. Computational experiments are performed to test the VNS algorithm against a benchmark problem set. The results show that the VNS algorithm is an effective approach for solving the MTFWS problem, capable of discovering many large-one frequent itemset with time-windows (FITW) with a larger time-coverage rate than the lower bounds, thus laying a good foundation for mining ARTW.

1. Introduction

The association rule (AR) is a relationship between data items. Association rule mining (ARM) is a class of computing techniques used to discover interesting relationships among a set of items by finding itemsets that frequently appear together in transactions. The *minsup–minconf* framework for ARM has been the most widely used ARM method since it was presented by Agrawal et al. [1] 20 years ago. There are two user-specified criteria, minimum support (*minsup*) and minimum confidence (*minconf*), that are used in the *minsup–minconf* framework. The term *minsup* specifies the minimum frequency of the itemsets in the transactions and the term *minconf* specifies the minimum confidence level of the associated relations between frequent itemsets. Agrawal and Srikant [2] decomposed the ARM problem into two sub-problems: (1) find all itemsets that have support above *minsup*, namely, the frequent itemsets; and (2) generate association rules from frequent itemsets that satisfy *minconf*. Because the first problem is very time-consuming, many subsequent research programs were focused on finding more efficient algorithms for discovering frequent itemsets. Agrawal and Srikant [2] presented a fast Apriori/AprioriTid algorithm that can generate all frequent itemsets by using the knowledge that all subsets of a frequent itemset are also frequent. There are thousands of research studies on association rule mining from various types of data; typical reviews can be found in, e.g., [3,4].

The time attribute of a transaction, i.e., the time when the transaction occurs, has attracted many researchers to discover knowledge regarding customers' purchasing patterns over time. These are sequential association rules (SARs). Agrawal and Srikant [5,6] first proposed an algorithm to mine the sequential patterns that locate the interdependencies, e.g., a specific order, among the occurrences of sequential events. Jea et al. [7] and Chen et al. [8] studied a hybrid sequential pattern in which the adjacent elements may be either consecutive or non-consecutive. Chen et al. [9,10] conducted further studies on time-interval sequential patterns and fuzzy multi-level sequential patterns. Kazienko [11] proposed a negative sequential pattern that indicates a certain set of items does not occur after a regular frequent sequence. Basha and Ameen [12] investigated the methodologies for the detection of a discord sub-sequence in a time series, especially with periodicity. Kim [24] incorporated a sequential navigation order to develop a streaming association rule for web data mining, and Peng and Liao [25] studied the problem of mining sequential patterns from multi-domain databases.

Another important utilization of the time attribute is to discover *temporal association rules* (TARs). In contrast to the traditional association rule, the temporal association rule adds a time constraint (it can be a point in time or a time range) to the rule to indicate when it holds. Chen and Petrounias [13] first addressed the temporal issues of association rules and outlined some research tasks. Ale and Rossi [14] studied TARs by utilizing the lifetime of an item/itemset, i.e., from the moment the item appeared to the

* Correspondence to: 37 Xueyuan Road, Haidian District. Beijing 100191, China. Tel.: +86 10 82316181; fax: +86 10 82328037.
E-mail addresses: qhzhao@buaa.edu.cn, qiuhongzhao@126.com (Q. Zhao).

0305-0548/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2013.09.018

moment it disappeared. Similar works can be found in [15] and in [16], where the individual time periods of different items have been used to discover the temporal association rules within the item/itemset's period. Verma et al. [17] used efficient T-tree and P-tree data structures to find calendar-based temporal association rules in time-dependent data. Chu et al. [18] and Lan et al. [19] studied mining rare temporal utility itemsets from temporal databases that appear infrequently in the current time window of large databases but are highly associated with specific data.

The *periodic association rule* (PAR), a special type of temporal association rule, has also attracted some attention in the research community. Current research has focused on finding the periodic purchasing activities occurring at regular time intervals. The PAR was first studied by Ozden et al. [20], with the goal of discovering the association rules that repeated themselves in every cycle with a fixed time span. Li et al. [21] presented a level-wise Apriori-based algorithm, named Temporal-Apriori, to discover the calendar-based periodic association rules, the periodicity of which is in the form of a calendar, e.g., day, week or month. Huang and Chang [22] investigated asynchronous partial periodic patterns in multi-event temporal databases. Lee et al. [23] relaxed the restriction of crisp periodicity, applying fuzzy periodicity, and have developed an algorithm for mining fuzzy periodic association rules. Xiao et al. [31] proposed a *minsup–minconf–minwin* framework for a mining association rule with a time-window (ARTW) on temporal transaction databases. They proposed an Apriori-like algorithm, named TW-Apriori, for mining ARTWs from real-time transaction databases.

In this study, we address a problem rising from the process of mining the ARTW, the maximum frequent time-window selection (MFTWS). We formulate the solution as an integer programming (IP) model and develop a variable neighbourhood search (VNS) algorithm to solve it with near-optimal solutions. Computational experiments are performed to test the model and the VNS algorithm.

The rest of the paper is organized as follows. In Section 2 we describe the formulation of the mining association rule with time-windows (ARTW). In Section 3, we describe the problem of MFTWS and formulate it as an IP model. Some properties of the model are provided. In Section 4, we develop a VNS algorithm as the solution approach for the MFTWS problem. In Section 5, computational experiments are performed to examine the problem model and the VNS algorithm. Finally, in Section 6, we present the conclusions.

2. Mining association rule with time-windows

The following nomenclature is used to formulate the association rule with time-windows (ARTW):

- I set of all items
- X item set such that $X \subseteq I$
- D set of all transactions in a temporal database
- d a transaction such that $d \in D$
- T total time span of D
- W set of time-windows
- w a time-window $w=[t_s, t_e]$, indicating a continuous time interval starting at time t_s and ending at time t_e
- $|w|$ width of time-window
- D^w set of transactions occurring in w
- $|D^w|$ number of transactions in D^w
- $D(X)^w$ set of transactions containing X and occurring in w
- $|D(X)^w|$ number of transactions in $D(X)^w$
- $s\%$ user-specified minimal support (*minsup*)
- $c\%$ user-specified minimal confidence (*minconf*)
- ω user-specified minimal width of time-window (*minwin*)

Definition 1. Frequent itemset with time-windows (FITW): An FITW can be expressed as X^W subjected to the following constraints:

(1) $X \subseteq$ I,
(2) for any two w_i, $w_j \in W$, satisfy $w_i \cap w_j = \phi$, and
(3) for each $w \in W$, satisfies $(|D(X)^w|)/(|D^w|)100\% \geq s\%$.

It can be read as *the itemset X appears frequently over all of the non-overlapped continuous time intervals in W with a support greater than or equal to a user-specified minimal support, s%.*

Definition 2. Association rule with time-windows (ARTW): An ARTW is an implication rule of $X \Rightarrow^W Y$ that holds if and only if it satisfies the following constraints:

(1) $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \phi, W \neq \phi$,
(2) X^W and Y^W are FITWs with respect to the user-specified minimal support $s\%$,
(3) for each $w \in W$, satisfy $(|D(X \cup Y)^w|)/(|D(X)^w|)100\% \geq c\%$, and
(4) for each $w \in W$, satisfies $|w| \geq \omega$.

The definition of an ARTW uses three thresholds to determine if it holds or not, two traditional ones, the *minsup* and the *minconf*, and a new one, the *minwin*. The philosophy underlying the new threshold is that an ARTW that holds only in a very narrow time-window may be noise or not important enough for special attention from market management. In the special case, when the minwin is broadened to the entire time span of the temporal database, then the ARTW degrades to the traditional association rule. When the time-windows of an ARTW are spaced equidistantly and of equal width, it becomes a periodic association rule. Therefore, the ARTW is a more general form of association rule that includes time as a factor.

Three indexes are defined to characterize the strengths of an ARTW, the *mean support*, the *mean confidence*, and the *time-coverage rate*.

Definition 3. Mean support of an ARTW: The mean support, $\bar{s}\%$, of an ARTW $X \Rightarrow^W Y$ is defined as $\bar{s}\% = (\sum_{w \in W}|w|s)/(\sum_{w \in W}|w|)100\%$, where s is the support of X in time-window w and $w \in W$, i.e., $s = (|D(X)^w|)/(|D^w|)$, $\forall w \in W$.

Definition 4. Mean confidence of an ARTW: The mean confidence, $\bar{c}\%$, of an ARTW $X \Rightarrow^W Y$is defined as $\bar{c}\% = (\sum_{w \in W}|w|c_w)/\sum_{w \in W}|w|100\%$, where c_w is the confidence of $X \Rightarrow^w Y$, i.e., $c_w = (|D(X \cup Y)^w|)/|D(X)^w|$.

Based on the definitions, an ARTW always has a *mean support* greater than or equal to *minsup* and a *mean confidence* greater than or equal to *minconf*; whereas the reverse law does not exist—a rule with larger *means support* or *mean confidence* does not guarantee it being a ARTW.

Definition 5. Time-coverage rate of ARTW: The time-coverage rate, $tc\%$, characterizes how the strength of the ARTW $X \Rightarrow^W Y$ holds over a time horizon and is defined as follows:

$$tc\% = \frac{\sum_{w \in W}|w|}{|T|}100\%,$$

where $\sum_{w \in W}|w|$ is the total length of the time-windows of $X \Rightarrow^W Y$, and $|T|$ is the total time span of the temporal database. In the special case where $tc\% = 100\%$, the ARTW is a traditional and *full-time* association rule; if $tc\%$ is less than 100%, e.g., 80% or 50%, the ARTW is a *part-time* association rule. At all times, $tc\% \geq \omega/|T|$.