Engineering Applications of Artificial Intelligence 69 (2018) 1–23

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Mining non-redundant closed flexible periodic patterns

Sayma Akther [a], Md. Rezaul Karim [a], Md. Samiullah [a], Chowdhury Farhan Ahmed [a,b,*]

[a] Department of Computer Science and Engineering, University of Dhaka, Bangladesh
[b] ICube Laboratory, University of Strasbourg, France

ARTICLE INFO

Keywords:
Data mining
Time series databases
Periodic patterns
Closed periodic patterns
Flexible patterns
Non-redundant patterns

ABSTRACT

Mining periodic patterns from time series databases is needed to predict any future situation. Flexible pattern mining is a special kind of periodic pattern mining where intermediate events can be overlooked purposely. Mining such patterns from time series data is advantageous due to its capability of modeling various real life scenarios. The goal of mining closed flexible patterns is to avoid unnecessary flexible patterns but preserving the same information of a complete set of patterns. Though it has wide range of application domains, existing algorithms failed to mine closed flexible patterns without generating any false positive, i.e. non-closed and /or redundant patterns. In this paper, a new algorithm $NRCFP$ (Non-Redundant Closed Flexible Pattern) has been proposed that generates complete set of non-redundant closed flexible patterns in time series databases. Three pruning techniques- $BackScan$ (existing), $RangeScan$ (proposed) and $Column\text{-}pruning$ (proposed) have been applied to avoid generation of non-closed patterns, redundant flexible patterns and fictitious patterns. Proposed $NRCFP$ efficiently mines non-redundant closed flexible periodic patterns. The performance of our algorithm has been extensively analyzed using several real-life databases based on runtime and memory consumption and compared with existing state-of-the-arts approach to prove effectiveness of the algorithm with respect to required processing time and memory consumption. Some applications of our proposed algorithm in various real life domains are discussed.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Data mining refers to the process of mining knowledge from large volume of data. One of the most important tasks in data mining is pattern mining which includes frequent pattern mining (Agrawal and Srikant, 1994; Ahmed et al., 2012a,b, 2011; Han et al., 2000b; Duong et al., 2014), closed-frequent pattern mining (Kim et al., 2004; Pei et al., 2000; Zaki and Hsiao, 2005), episode mining (Mannila et al., 1995, 1997), event labeling (Fanaee-T and Gama, 2014), inter transaction pattern mining (Tung et al., 1999; Lu et al., 1998; Pasquier et al., 1999a), graph pattern mining (Yan and Han, 2002; Samiullah et al., 2014; Fariha et al., 2015), sequential pattern mining (Srikant and Agrawal, 1996b; Agrawal and Srikant, 1995; Pei et al., 2004; Wang and Han, 2004), periodic pattern mining (Rasheed et al., 2011), association rule mining (Han and Kamber, 2000) etc.

Frequent patterns can provide critical information to the decision makers from the collected data in solving problems for various domains, such as marketing, medical science, meteorology, finance etc. Periodic pattern is a collection of data values, gathered generally at uniform interval of time, reflecting certain behavior of an entity (Chanda et al., 2015). In real-life, time series data mining techniques are mostly used to mine interesting knowledge from the recurring events in the databases (Han and Kamber, 2000; chung Fu, 2011). Mining of periodic pattern is performed on time series database (Rasheed et al., 2011; Rasheed and Alhajj, 2010).

Now-a-days, flexible periodic pattern mining is an interesting problem where a periodic pattern is called flexible if some intermediate events between any pair of events are less significant (do not care events) and the number of such events is not fixed.

A flexible periodic pattern is considered to be closed if there exists no super-pattern having the same support count, where support is an indication of how frequently the pattern appears in the database, i.e. percentage of sequences that contain the pattern. It avoids unnecessary flexible periodic patterns without losing any information. Therefore mining closed flexible periodic pattern is an interesting problem.

One of the recent and most efficient state-of-the-arts approach (Wu and Lee, 2010) generates closed flexible periodic patterns of different

* Corresponding author at: Department of Computer Science and Engineering, University of Dhaka, Bangladesh.
E-mail addresses: holy.sayma@gmail.com (S. Akther), rkarim@cse.univdhaka.edu (M. Rezaul Karim), samiullah@cse.univdhaka.edu (M. Samiullah), cfahmed@unistra.fr, farhan@cse.univdhaka.edu (C.F. Ahmed).

https://doi.org/10.1016/j.engappai.2017.11.005
Received 6 August 2016; Received in revised form 12 July 2017; Accepted 7 November 2017
0952-1976/© 2017 Elsevier Ltd. All rights reserved.

Table 1
Sample time series database of a patient health monitoring system.

Day	Time	Test code	Explanation	Result
04–21–2015	7.09	58	Sugar test(Breakfast)	130(High)
04–21–2015	9.09	23	Blood pressure	069(Low)
04–21–2015	13.08	60	Sugar test(Lunch)	052(Low)
04–21–2015	17.08	24	Cholesterol test	212(Very high)
04–21–2015	20.51	62	Sugar test(Dinner)	031(Bad)
04–21–2015	23.00	23	Blood pressure	040(Very low)
04–22–2015	7.35	58	Sugar test(Breakfast)	076(Low)
04–22–2015	13.40	60	Sugar test(Lunch)	047(Bad)
04–22–2015	20.16	62	Sugar test(Dinner)	157(High)
04–22–2015	21.36	24	Cholesterol test	041(Bad)
04–22–2015	22.55	23	Blood pressure	097(Normal)
04–22–2015	23.40	23	Blood pressure	107(Normal)
04–23–2015	6.33	23	Blood pressure	175(High)
04–23–2015	7.25	58	Sugar test(Breakfast)	051(Low)
04–23–2015	13.35	60	Sugar test(Lunch)	012(Bad)
04–23–2015	20.44	62	Sugar test(Dinner)	167(High)
04–23–2015	22.52	24	Cholesterol test	030(Bad)
04–23–2015	23.12	23	Blood pressure	089(Normal)
04–24–2015	7.46	58	Sugar test(Breakfast)	075(Low)
04–24–2015	13.25	60	Sugar test(Lunch)	011(Bad)
04–24–2015	15.35	23	Blood pressure	220(Very high)
04–24–2015	20.25	62	Sugar test(Dinner)	117(High)
04–24–2015	22.33	24	Cholesterol test	059(Low)
04–24–2015	23.26	23	Blood pressure	031(Very low)

length incrementally until no more patterns can be generated. The algorithm generates closed flexible periodic patterns, but have some remarkable limitations. One of the limitation is generation of redundant closed and non-closed flexible patterns. It is a significant drawback to create non-closed patterns simultaneously with closed flexible patterns. The algorithm also consumes huge amount of memory due to creating large number of non-closed patterns. However, it uses some pruning techniques to remove redundant patterns and non-closed patterns. A post pruning technique is applied after generation of redundant pattern, hence the cost involved in generation of redundant patterns cannot be avoided. To identify non-closed patterns, the checking of superset has been applied after the pattern generation. It creates an index table which creates worthless pattern due to containing of useless fields in index table. These drawbacks of the existing algorithm motivated us to introduce a better efficient algorithm $NRCFP$ (**N**on-redundant **C**losed **F**lexible **P**eriodic pattern) (NRCFP) with some interesting efficient rules to find non-redundant closed flexible periodic patterns over the existing one.

1.1. Motivating example

As a motivating real life example, consider an insulin deficient patient health monitoring status for different days where different types of tests are repeatedly performed every day in the same time period. Date, time, test code, explanation of every test codes and results of every tests are stored in a time series database shown in Table 1. (The dataset constructed in this table is used for all illustrative discussions and examples unless stated otherwise.) Suppose a physician want to predict the future health condition of that patient by finding a periodic pattern from the time series database as shown in Table 1 that reflect patient condition. According to Table 1, some flexible periodic patterns are found such as $\{Low[LSE]Bad\}$, $\{Low[LSE]High\}$, $\{Bad[LSE]High\}$, $\{High[LSE]Low\}$ and $\{Low[LSE]Bad[LSE]High\}$ where *LSE* means one less significant event and $\{[LSE]\}$ means more than one less significant events. Assume that a diabetic specialist is interested in only glucose measurement test and considers other intermediate test events as less significant, since the patient is an insulin deficient patient. From the pattern $\{Low[LSE]Bad[LSE]High\}$, diabetic specialists can predict meaningful future condition of insulin deficient patients.

Suppose, if a physician sets *minimum support threshold* to 50%, then some patterns, i.e., $\{Low[LSE]High\}$, $\{Bad[LSE]High\}$ and $\{Low[LSE]Bad[LSE]High\}$ are mined, having support 75%. Patterns $\{Low[LSE]High\}$ and $\{Bad[LSE]High\}$ are the subset of the pattern $\{Low(LSE)Bad(LSE)High\}$. So they are not closed patterns. However, the pattern $\{High[LSE]Low\}$ and $\{Low[LSE]Bad[LSE]High\}$ are closed patterns because they have no superset with same support. However, pattern $\{Low[LSE]Bad\}$ remains in the pattern $\{Low[LSE]Bad[LSE]High\}$ but having different support 100%. So both the patterns are closed pattern. The closed flexible patterns $\{Low[LSE]Bad\}$, $\{High[LSE]Low\}$ and $\{Low[LSE]Bad[LSE]High\}$ are generated by avoiding unnecessary flexible patterns but preserving the same information of a complete set of flexible patterns.

Hence, huge amount of related and similar patterns are compressed which leads to faster processing and memory efficient approach to mine lossless and similar number of association rules. It is already known that, mining closed flexible periodic patterns in time series databases which is memory and time efficient. Furthermore, it will be a better solution if unnecessary, repetitive and redundant patterns can be avoided. Let us give an example to understand clearly the necessity of avoiding the generation of redundant patterns. Suppose $\{Low[LSE]High\}$ is a pattern where $[LSE]$ represents 1 or 2 less significant events, i.e., $\{Low, LSE, High\}, \{Low, LSE, LSE, High\}$. $\{Low[LSE]High\}$ is a redundant pattern with respect to pattern $\{Low, LSE(1), Bad, LSE(2), High\}$ where $LSE(1)$ and $LSE(2)$ represent 0 or 1 less significant event, i.e., $(\{Low, Bad, High\}, \{Low, LSE, Bad, High\}$, $\{Low, Bad, LSE, High\}$, $\{Low, LSE, Bad, LSE, High\})$. So, mining of the pattern $\{Low[LSE]High\}$ consumes extra memory and wastes time.

1.2. Contribution

In this paper, a new and efficient approach has been proposed to mine non-redundant closed flexible periodic patterns. An efficient algorithm, $NRCFP$ (**N**on-redundant **C**losed **F**lexible **P**eriodic pattern) has been designed for mining efficiently the complete set of closed flexible periodic. The key contributions of this research work are describes as follows:

- Development of an efficient non-redundant algorithm $NRCFP$ to mine non-redundant closed flexible periodic patterns from time series database.
- Introduction of a new pruning technique *Range Scan* which is applied to stop generating non-closed and redundant flexible patterns when mine the closed flexible periodic patterns.
- Proposal of another memory efficient new pruning technique is called *column pruning* that optimize index table to avoid generating fictitious patterns.
- Effective utilization of existing Back Scan technique is done to eradicate non-closed flexible periodic patterns.
- Significant improvement of performance by avoiding the generation of non-closed patterns, redundant patterns and fictitious patterns.
- Extensive performance analysis and comparison is performed using several benchmark and real life datasets with existing closed flexible pattern mining approach to recognize *NRCFP* as a scalable and efficient lattice mining approach.
- Presentation of several applications using non-redundant closed flexible periodic patterns.

The rest of the paper is organized as follows. Section 2 provides the preliminary definition and related work. Section 3 describes the complete methodology of our idea and contains detailed explanation of our proposed algorithm. The properties of our algorithm is illustrated in Section 4. In Section 5, extensive experiments have been presented on the real-life datasets. Section 6 provides some real life scenarios where *NRCFP* can be proved useful. Finally Section 7 concludes our paper with future research direction.

Download English Version:

https://daneshyari.com/en/article/6854244

Download Persian Version:

https://daneshyari.com/article/6854244

Daneshyari.com