Information Sciences 000 (2018) 1–18

Contents lists available at ScienceDirect

Information Sciences

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

Frequent pattern discovery with tri-partition alphabets

Fan Min[a,*], Zhi-Heng Zhang[b,a], Wen-Jie Zhai[a], Rong-Ping Shen[a]

[a] School of Computer Science, Southwest Petroleum University, Chengdu 610500, China
[b] School of Science, Southwest Petroleum University, Chengdu 610500, China

ARTICLE INFO

Article history:
Received 2 August 2017
Revised 16 March 2018
Accepted 3 April 2018
Available online xxx

Keywords:
Pattern discovery
Sequence
Tri-partition
Tri-pattern

ABSTRACT

The concept of patterns is the basis of sequence analysis. There are various pattern definitions for biological data, texts, and time series. Inspired by the methodology of three-way decisions and protein tri-partition, this paper proposes a frequent pattern discovery algorithm for a new type of pattern by dividing the alphabet into strong, medium, and weak parts. The new type, called a tri-pattern, is more general and flexible than existing ones and is therefore more interesting in applications. Experiments were undertaken on data in various fields to reveal the universality of this new pattern. These include protein sequence mining, petroleum production time series analysis, and forged Chinese text keyword mining. The results show that tri-patterns are more meaningful and desirable than the existing four types of patterns. This study enriches the semantics of sequential pattern discovery and the application fields of three-way decisions.

1. Introduction

Pattern discovery has been widely studied in machine learning tasks such as image recognition [12] and text analysis [34]. A sequential pattern is a (totally or partially) ordered subsequence of a transactional dataset [1], symbolic sequence [29], numeric time series [26], or other data sequence. Sequential pattern discovery (also called pattern mining) [1,38] refers to the discovery of all frequent sequential patterns from these data. For transactional datasets such as retailing/market-baskets and planning [11], a pattern is a sequence of events, where each event is represented by an itemset. This type of representation has horizontal and vertical scalability, leading to general association rules. For symbolic sequences such as protein [24] and text, a pattern is a subsequence of characters, also called a string. It can easily represent codons or keywords [23]. For numeric time series such as stock prices [25] and petroleum production [26], where the trend of fluctuation is essential to the stock holders and petroleum cooperation, a pattern is a subsequence of real values.

One interesting research direction of symbolic sequential pattern discovery involves the generalization of patterns. A plain pattern is a subsequence that must be exactly matched. A wildcard [3], which is also called a motif [39], matches any character in the alphabet. A wildcard gap [3] matches any subsequence within the length constraint. In this way, a pattern with wildcard gaps is able to handle noise or shifts [41]. To enrich the semantics of the pattern, the alphabet is divided into weak and strong parts. A weak-wildcard gap [32] matches a subsequence of weak characters. Consequently, a pattern with weak-wildcard gaps [32] not only ignores weak characters, but also maintains strong ones.

In this paper, we introduce a more general and flexible type of pattern called a tri-pattern that is inspired by three-way decisions (3WD) [43]. The alphabet is partitioned into three parts Γ, Λ, and Ω, corresponding to strong, medium,

* Corresponding author.
E-mail address: minfanphd@163.com (F. Min).

https://doi.org/10.1016/j.ins.2018.04.013
0020-0255/

and weak characters, respectively. In this situation, a tri-wildcard gap $G = (N, M)$ matches any sequence of medium/weak characters with a length between N and M. A tri-pattern is a sequence of strong/medium characters containing periodic tri-wildcard gaps, where the term "periodic" indicates that the gaps between any two adjacent characters are identical. The idea of a tri-partition has been widely adopted in many applications. Biologists partition the human amino acid alphabet into essential, conditional, and nonessential amino acids. Petroleum experts partition oil production fluctuation into significant, insignificant, and minor changes. The tri-partition of situations and actions [16,20,42] also has attracted a large amount of research interest in data mining, especially 3WD [6,10,30]. To the best of our knowledge, however, a tri-partition of an alphabet has not been considered in the 3WD research community.

Tri-patterns are more general than the four existing types of patterns. Let tri-patterns be Type I patterns, plain patterns be Type II patterns, patterns with periodic wildcard gaps [3] be Type III patterns, patterns with periodic weak-wildcard gaps [32] be Type IV patterns, and strong patterns with periodic weak-wildcard gaps [32] be Type V patterns. Types II through V are the special cases of Type I where $\Lambda = \Omega = \emptyset$, $\Lambda = \Sigma$, $\Omega = \emptyset$, and $\Lambda = \emptyset$, respectively. Similar to existing frequent pattern discovery problems, a new problem for Type I patterns is defined. Using the Apriori property of the new problem, an Apriori algorithm is designed for efficient pattern discovery and tree pruning.

We compare these five types of patterns in three application areas to reveal the universality of the new pattern. The first application is human protein sequence mining. Essential, conditional, and nonessential amino acids correspond to strong, medium, and weak characters, respectively. Twenty sequences were concatenated to discover frequent patterns. These patterns and their popularity in the original sequences were analyzed. Tri-patterns are the most meaningful type, while the other types, especially Type II, suffer from too many weak characters. For the pattern popularity, tri-patterns have the best minimal and second-best average performance.

The second application is oil well daily production time series analysis. A coding table was designed to convert the numeric time series into a nominal sequence. Significant, insignificant, and minor changes correspond to strong, medium, and weak characters, respectively. We used two years of well data to discover frequent patterns, which are finally matched in the original time series. Compared with distance-based approaches that analyze numeric time series directly, the coding and mining approach handles different levels of fluctuations more easily. Observation on the original time series shows that tri-patterns match some similar subsequences with minor differences. In contrast, the four existing types of patterns either exclude some similar subsequences or include dissimilar ones.

The third application is forged Chinese text mining. Notional words, function words, and special characters correspond to strong, medium, and weak characters, respectively. Four sets of text for news, novels, history, and law were collected. Forged Chinese text occurs when some text creators distribute advertisements for illegal items such as forged money or invoices. They may insert some characters into the text that makes the text still readable, but difficult for automated analysis. In this way, the text can avoid spam filters and similar blockers. Our purpose is to extract keywords from the forged text that are also keywords in the original text. This problem is closely related to fuzzy text matching and keyword extraction. With tri-patterns, the algorithm achieves the highest accuracy in mining keywords from forged text.

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 defines the new types of pattern, analyzes the relationships among the five types, and presents the new pattern discovery problem. Section 4 demonstrates the Apriori property and proposes the new algorithm. Section 5 explains the experimental settings and results on three kinds of real-world data. Finally, the concluding remarks are discussed in Section 6.

2. Related work

In this section, we present some related work concerning 3WD and sequential pattern discovery. The formal definition of some existing concepts is presented in the next section.

2.1. Three-way decision

In rough set theory [27], using to an equivalence relation, the universe is partitioned into three disjoint regions. The positive, negative, or boundary regions are the set of objects that are definitely, definitely not, or possibly members of the target set, respectively. However, this strict division often hinders its application in practice. Probabilistic rough set models such as decision-theoretical rough sets (DTRS) [44] and variable precision rough sets [50] were introduced to address this issue. Among them, DTRS is a sound theory based on Bayesian decision making. The required parameters are systematically determined based on the costs of various decisions [43,44].

3WD has substantially advanced by becoming a methodology of divide and conquer [43] rather than a concrete technique. One task of this methodology is to construct a tri-section or tri-partition of the universal set. The other task is to act upon objects in one or more regions by developing appropriate strategies. 3WD is a class of effective methods and heuristics commonly used in human problem solving and information processing.

Recently, various theories have been inspired by 3WD. Three-way formal concept analysis [28] is constructed using the three-way operators and their inverses. Three-way cognition computing [16] focuses on concept learning via multi-granularity from the viewpoint of cognition. The three-way fuzzy sets method [4] constructs a three-way, three-valued, or three-region approximation with a pair of thresholds on the fuzzy membership function. Three-way decisions space [9] uni-

Please cite this article as: F. Min et al., Frequent pattern discovery with tri-partition alphabets, Information Sciences (2018), https://doi.org/10.1016/j.ins.2018.04.013