Information Sciences 222 (2013) 126–146

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Discovery of fuzzy quantitative sequential patterns with multiple minimum supports and adjustable membership functions

Tony Cheng-Kui Huang *

Department of Business Administration, National Chung Cheng University, 168, University Rd., Min-Hsiung, Chia-Yi, Taiwan, ROC

ARTICLE INFO

Article history:
Received 27 April 2009
Received in revised form 22 July 2012
Accepted 24 July 2012
Available online 10 August 2012

Keywords:
Data mining
Sequential pattern
Multiple minimum support
Adjustable membership function
Quantitative data

ABSTRACT

The objective of mining quantitative sequential patterns is to discover complete sets of sequential patterns with quantities in databases. Although this novel type of pattern considers information, compared to traditional sequential patterns, it contains a sharp boundary problem; that is, when the quantity of an item is near the boundary of two predetermined quantity intervals, it can be either ignored or overemphasized. Hong et al. proposed a new type of pattern, called a fuzzy quantitative sequential pattern (FQSP), to solve this problem. Although this type of pattern provides summary information to expedite decision making, FQSPs confront unrealistic circumstances, that is, they offer only a unique minimum support (min_sup) to all items for the whole database. A higher min_sup fails to find rare items with higher profit, whereas a lower min_sup leads to a combinatorial explosion problem. Instead of this type of mechanism, we used the idea of multiple minimum supports to mine an FQSP. Furthermore, the utility of FQSP is also disturbed because this approach uses only a single membership function without considering the price–quantity relation. The idea of adjustable membership functions was also provided to address this problem. This study proposes a new model to discover an FQSP with both multiple minimum supports and adjustable membership functions. Experiments using synthetic and real datasets demonstrated the computational efficiency, scalability, and effectiveness of the model.

1. Introduction

Sequential pattern mining was first introduced in the mid-1990s to discover the ordinal relations that are existed among items [2]. As one of the crucial data mining approaches [4,5,9,11], sequential pattern mining has attracted considerable attention from both academic researchers and practitioners.

An example of a sequential pattern is when a customer returns to purchase a printer toner and printer paper after purchasing a color laser printer. To further discover the quantities associated with items, a number of recent studies have used quantitative association rules [26] or quantitative sequential patterns [13] in their proposed solutions. An example of a quantitative sequential pattern is described as follows:

> After purchasing "1" color laser printer at a computer store, a customer returns to purchase "2–4" printer toner cartridges and "8–10" packs of printer paper.

* Tel.: +886 5 2720411x34319; fax: +886 5 2720564.
E-mail address: bmahck@ccu.edu.tw

0020-0255/$ - see front matter
http://dx.doi.org/10.1016/j.ins.2012.07.047

This quantitative sequential pattern shows both the order of items and the associated ranges of the quantities. This type of quantitative pattern is more informative than traditional sequential patterns because it shows the items that were purchased and the quantities of the purchased items.

In this paper, we introduce a new sequential pattern mining model that combines the *fuzzy sets* and the idea of *multiple minimum supports* to provide more flexibility and information to the result.

1.1. Mining fuzzy patterns

Mining quantitative sequential patterns shows the quantity intervals of items. However, the proposed approach may have a *sharp boundary problem* [3]. In other words, when a quantity is close to the boundary of two adjacent quantity intervals, it can be either ignored or overemphasized. This problem can be adequately addressed by using a *fuzzy partition method*, allowing this quantity to be 50% in one interval and 50% in another interval simultaneously. This example shows that the fuzzy partition method provides smooth transition between members and non-members of a set, making the result more interpretative than classical ones.

According to this idea, Hong et al. [7] proposed the fuzzy quantitative sequential pattern (FQSP) to address the problem. They used fuzzy terms to represent quantities of items in a pattern. Membership functions can be assigned to represent fuzzy terms in place of a range of quantities. Chen and Huang [3] furthered the concept by developing a *divide-and-conquer approach* to improve the effectiveness of FQSP to make it more practical for practitioners.

In summary, FQSPs manage the sharp boundary problem, and also assist managers to expedite their decision making by using more natural linguistic terms.

1.2. The unrealistic problem

Although mining FQSPs offer more straightforward information for managers, there are two concerns remining to be solved. We describe them as follows.

1.2.1. Multiple minimum supports

One concern is that FQSPs define only one fixed minimum support (min_sup) to all items for the whole database, thereby hindering its practicability in the real world. We outline the FQSP procedures to give preliminary explanation in the next paragraph.

With a set of sequences, where each sequence consists of a list of itemsets and each itemset consists of a set of items with quantities, and with a user-specified min_sup threshold, FQSP mining finds all the frequent subsequences with fuzzy terms, the occurrence frequency of which in the set of sequences is greater than min_sup. The main element that makes FQSP mining practical is the threshold, min_sup. It is used to prune the search space and limit the number of generated rules. However, using only one fixed min_sup value implies that all items are of the same nature or of similar frequencies in the database. This does not fit in the complicated cases in real life.

Consumers usually purchase durable or expensive items less frequently than expendable and cheap items. Considering the previous example, the demand of color laser printers is lower than the demand of printer toners or printer paper. Therefore, frequent purchases of this type of item are unlikely to happen. In such cases, if a high value of min_sup is given, it would fail to discover the patterns that involve infrequent items or rare items. By contrast, if the given value of min_sup is low, it would acquire excessive meaningless knowledge for decision markers. This approach may also cause combinatorial explosion, leading to deteriorated performance.

This dilemma is called the *rare item problem* [22]. To solve this problem, Liu et al. [15] proposed an approach to mine association rules so that different min_sups can be assigned to different items. This approach enables users to produce rare item rules without the combinatorial explosion problem. According to the work of Liu et al. [15], Lin [17] extended the idea to discover traditional sequential patterns by a PrefixSpan-based algorithm. However, previous studies did not consider quantitative data in the sequential mining model.

1.2.2. Adjustable membership functions

This study proposes a novel algorithm to discover FQSPs with multiple minimum supports (FQSP-MMS). The FQSP-MMS algorithm is the extension of Generalized Sequential Pattern (GSP), which uses a stage-by-stage process for generating frequent patterns [25].

Although we adopt the MMS idea from the classical partition case, the FQSP model in the fuzzy case remains obstructed because it uses only one membership function without considering the *price–quantity relation*. For example, the implication of a pattern "*Color Laser Printer with Low quantity*" must be distinguished from that of another pattern "*Printer Toner with Low quantity*" although both patterns are assigned with a same fuzzy term. Managers may specify different definitions of *Low* quantity for *Color Laser Printers* and *Printer Toner*. Items with different prices result in different quantity demands; therefore, different membership functions must be dispatched to calculate their fuzzy term supports.