Expert Systems with Applications 37 (2010) 4560–4567

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Linguistic data mining with fuzzy FP-trees ☆

Chun-Wei Lin [a], Tzung-Pei Hong [b,c,*], Wen-Hsiang Lu [a]

[a] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, ROC
[c] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan, ROC

ARTICLE INFO

Keywords:
Fuzzy data mining
Fuzzy-set
Quantitative value
Fuzzy FP-trees
Frequent fuzzy patterns

ABSTRACT

Due to the increasing occurrence of very large databases, mining useful information and knowledge from transactions is evolving into an important research area. In the past, many algorithms were proposed for mining association rules, most of which were based on items with binary values. Transactions with quantitative values are, however, commonly seen in real-world applications. In this paper, the frequent fuzzy pattern tree (fuzzy FP-tree) is proposed for extracting frequent fuzzy itemsets from the transactions with quantitative values. When extending the FP-tree to handle fuzzy data, the processing becomes much more complex than the original since fuzzy intersection in each transaction has to be handled. The fuzzy FP-tree construction algorithm is thus designed, and the mining process based on the tree is presented. Experimental results on three different numbers of fuzzy regions also show the performance of the proposed approach.

1. Introduction

Years of effort in data mining have produced a variety of efficient and effective techniques. Depending on the classes of the knowledge derived, the mining approaches may be classified as finding association rules, classification rules, clustering rules and sequential patterns (Agrawal & Srikant, 1995), among others. Especially, finding association rules in transaction databases is most commonly seen in data mining (Agrawal, Imielinski, & Swami, 1993a; Agrawal, Imielinski, & Swami, 1993b; Agrawal & Srikant, 1994; Chen, Han, & Yu, 1996; Cheung, Lee, & Kao, 1997).

In the past, many algorithms for mining association rules from transactions were proposed. Most of the approaches were based on the *Apriori* algorithm (Agrawal et al., 1993a), which generated and tested candidate itemsets level by level. This may cause iterative database scans and high computational costs. Han et al. thus proposed the Frequent-Pattern-tree (FP-tree) structure for efficiently mining association rules without generation of candidate itemsets (Han, Pei, & Yin, 2000). The FP-tree was used to compress a database into a tree structure which stored only large items. It was condensed and complete for finding all the frequent patterns. The construction process was executed tuple by tuple, from the first transaction to the last one. After that, a recursive mining procedure called FP-growth was executed to derive frequent patterns from the FP-tree.

In these years, the fuzzy-set theory (Zadeh, 1965) has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning (Kandel, 1992). Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains (Hong & Chen, 1999, 2000). As to fuzzy data mining, several approaches have been proposed. For example, Hong et al. proposed a fuzzy mining algorithm for managing quantitative data (Hong, Kuo, & Chi, 1999b). It was based on the Apriori algorithm. Basically, it first used membership functions to transform each quantitative value into a fuzzy set in linguistic terms. It then calculated the cardinality of each linguistic term on all the transaction data. The mining process based on the cardinalities was then performed to find linguistic frequent itemsets and association rules.

Papadimitriou et al. proposed an approach based on FP-trees to find fuzzy association rules (Papadimitriou & Mavroudi, 2005). In their approach, each item in the transactions was transferred into only two fuzzy regions with individual fuzzy values. A threshold was set and a fuzzy region in a transaction would be removed if its fuzzy value was smaller than the threshold. In this process, only the local frequent fuzzy 1-itemsets kept in each transaction were used for mining. The fuzzy regions which were close to but below the threshold would provide no contribution at all to the mining. Thus, some fuzzy regions would not be frequent even the summation of its fuzzy values in the database was larger than or equal to

☆ This is a modified and expanded version of the paper "Mining fuzzy association rules based on fuzzy FP-trees", presented at The 16th National Conference on Fuzzy Theory and its Applications, Taiwan, 2008.

* Corresponding author. Address: Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, ROC.
E-mail addresses: p7895122@mail.ncku.edu.tw (C.-W. Lin), tphong@nuk.edu.tw (T.-P. Hong), whlu@mail.ncku.edu.tw (W.-H. Lu).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.12.052

the minimum support. Besides, the expression of fuzzy patterns with more fuzzy regions was straight. The approach did not use any fuzzy operation to combine fuzzy regions together. It made the mined fuzzy rules a little hard to understand.

In this paper, we attempt to extend the FP-tree mining process for handling quantitative data from the global values of fuzzy regions. A new fuzzy FP-tree is thus proposed, which is a data structure keeping frequent fuzzy regions. Besides, fuzzy operations are considered in forming itemsets with more than one fuzzy region. For achieving this purpose, the proposed fuzzy FP-tree is a little more complex than the original one and than that proposed by Papadimitriou and Mavroudi (2005). Based on the proposed approach, the frequent itemsets are efficiently expressed and mined out in linguistic terms, which are more natural and understandable for human beings.

The remainder of this paper is organized as follows. Related works are reviewed in Section 2. The notation used in the algorithm is explained in Section 3. The proposed fuzzy FP-tree construction algorithm is described in Section 4. An example to illustrate the proposed algorithm is given in Section 5. Experimental results for showing the performance of the proposed algorithm are provided in Section 6. Conclusions are finally given in Section 7.

2. Review of related works

In this section, some related researches are briefly reviewed. They are fuzzy-set concepts, mining association rules from quantitative data, and FP-tree algorithm.

2.1. Fuzzy-set concepts

A fuzzy set is an extension of a crisp set. Crisp sets allow only full membership or no membership at all, whereas fuzzy sets allow partial membership. Besides, each element may belong to more than one set. In a crisp set, the membership of an element x in set A is described by a characteristic function $u_A(x)$, where

$$u_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

The fuzzy-set theory extends this concept by defining partial memberships, which can take values ranging from 0 to 1. A membership function is formally defined as follows (Kosko, 1997; Zadeh, 1965):

$$u_A : X \to [0,1],$$

where X refers to the universal set for a specific problem. Assuming that A and B are two fuzzy sets with membership functions $u_A(x)$ and $u_B(x)$, respectively. The following common fuzzy operators can be defined as follows:

(1) The intersection operator:

$$u_{A\cap B}(x) = u_A(x)\ \tau\ u_B(x),$$

where τ is a *t-norm* operator. That is, τ is a function of $[0,1]*[0,1] \to [0,1]$ and must satisfy the following conditions for each $a,b,c \in [0,1]$:

(i) $a\ \tau\ 1 = a$;
(ii) $a\ \tau\ b = b\ \tau\ a$;
(iii) $a\ \tau\ b \geqslant c\ \tau\ d$ if $a \geqslant c, b \geqslant d$;
(iv) $a\ \tau\ b\ \tau\ c = a\ \tau\ (b\ \tau\ c) = (a\ \tau\ b)\ \tau\ c$.

Two instances of a *t-norm* operator for $a\ \tau\ b$ are $min(a, b)$ and $a * b$.

(2) The union operator:

$$u_{A\cap B}(x) = u_A(x)\ \rho\ u_B(x),$$

where ρ is an *s-norm* operator. That is, ρ is a function of $[0,1]*[0,1] \to [0,1]$ and must satisfy the following conditions for each $a,b,c \in [0,1]$:

(i) $a\ \rho\ 0 = a$;
(ii) $a\ \rho\ b = b\ \rho\ a$;
(iii) $a\ \rho\ b \geqslant c\ \rho\ d$ if $a \geqslant c, b \geqslant d$;
(iv) $a\ \rho\ b\ \rho\ c = a\ \rho\ (b\ \rho\ c) = (a\ \rho\ b)\ \rho\ c$.

Two instances of an *s-norm* operator for $a\ \rho\ b$ are $max(a,b)$ and $a + b - a * b$.

(3) The α -cut operator:

$$A_\alpha(x) = \{x \in X | u_A(x) \geqslant \alpha\},$$

where A_α is an α-*cut* of a fuzzy set A. A_α thus contains all the elements in the universal set X that have membership grades in A greater than or equal to the specified value of α. These fuzzy operators will be used in our fuzzy FP-tree mining algorithm to derive fuzzy association rules.

2.2. Mining algorithms for fuzzy association rules

It is useful to extract knowledge via data from the real world and to represent it in a comprehensible form. Linguistic representation is popular and may help knowledge more understandable to human beings. It is also easily implemented by fuzzy sets, since the fuzzy-set theory is concerned with quantifying and reasoning using natural language. Several fuzzy mining approaches have been proposed to find interesting linguistic association rules or sequential patterns from transaction data with quantitative values.

For example, Chan et al. proposed an F-APACS algorithm to mine fuzzy association rules (Chan & Au, 1997). They first transformed quantitative attribute values into linguistic terms and then used the adjusted difference analysis to find interesting associations among attributes. Kuok et al. proposed a fuzzy mining approach to handle numerical data in databases and to derive fuzzy association rules (Kuok, Fu, & Wong, 1998). At nearly the same time, Hong et al. proposed a fuzzy mining algorithm to mine fuzzy rules from quantitative transaction data (Hong & Chen, 1999, 2000; Hong et al., 1999b; Hong, Kuo, & Chi, 1999a). Besides, many mining methods for finding fuzzy association rules have also been proposed (Kaya & Alhajj, 2004; Shen, Wang, & Yang, 2004; Srikant & Agrawal, 1996; Yue, Tsand, Yeung, & Shi, 2000), and some related researches are still in progress.

2.3. The FP-tree mining algorithm

Han et al. proposed the Frequent-Pattern-tree structure (FP-tree) for efficiently mining association rules without generation of candidate itemsets (Han et al., 2000). The FP-tree mining algorithm consists of two phases. The first phase focuses on constructing the FP-tree from a database, and the second phase focuses on deriving frequent patterns from the FP-tree. Three steps are involved in FP-tree construction. The database is first scanned to find all items with their counts. The items with their supports equal to or larger than a predefined minimum support are selected as frequent 1-itemsets (items). Next, the frequent items are sorted in descending frequency. At last, the database is scanned again to construct the FP tree according to the sorted order of frequent items. The construction process is executed tuple by tuple, from the first transaction to the last one. After all transactions are processed, the FP tree is completely constructed.

After the FP tree is constructed from a database, a mining procedure called FP-growth (Han et al., 2000) is executed to find all frequent itemsets. FP-growth does not need to generate candidate itemsets for mining, but derives frequent patterns directly from the FP tree. A conditional FP tree is generated for each frequent item,