ATLANTIS PRESS

*Advances in Intelligent Systems Research, volume 131*

International Conference of Organizational Innovation (ICOI 2017)

# Using Fuzzy FP-Growth for Mining Association Rules

Chien-Hua Wang[*1]
School of Management
Fujian University of Technology
Fujian Province, 350118, China
[*]Corresponding author: wangthuck@gmail.com

Li Zheng[2]
School of Management
Fujian University of Technology
Fujian Province, 350118, China
zzhenglimail@gmail.com

Xuelian Yu[3]
School of Management
Fujian University of Technology
Fujian Province, 350118, China
Yuxuelian_2010@163.com

XiDuan Zheng[4]
School of Management
Fujian University of Technology
Fujian Province, 350118, China
476801980@qq.com

***Abstract*—This paper aims to use fuzzy set theory and FP-growth derived from fuzzy association rules. At first, we apply fuzzy partition method and decide a membership function of quantitative value for each transaction item. Next, we implement FP-growth to deal with the process of data mining. In addition, in order to understand the impact of fuzzy FP-growth algorithm and other fuzzy data mining algorithms on the execution time and the numbers of generated association rule, the experiment will be performed by using different thresholds. Lastly, the experiment results show fuzzy FP-growth algorithm is more efficient than other existing methods.**

***Keywords—data mining; fuzzy association rule; FP-growth***

## I. INTRODUCTION

Data mining is a methodology for the extraction of new knowledge from data. This knowledge may relate to problems that we want to solve [12]. Thus, data mining can ease the knowledge acquisition bottleneck in building prototype systems [4]. If data mining extracting can effectively be applied on all varieties of analysis, the process of decision-making in business will be fairly smooth.

In common transactions, the association rule (X→Y) is the most popular mean. The purpose is to search for the relation that exists among items of database. The relation reflects that exists (X) appear, other items(Y) are likely to appear as well [2]. For instance, when a customer purchases bread, one might also get milk along with it. Thus association rules can assist decision makers to scope out the possible items that are likely to be purchased by consumers. Meanwhile, it facilitates planning marketing strategies [1].

This paper mainly focuses on the association rule technique. In the conventional association rule algorithm, scanning database takes enormous time particularly when one uses Apriori algorithm, which often affects the efficiency in data mining. To solve the drawback aforementioned, Han et al. [3] proposed a mining method, called Frequent-Pattern growth (FP-growth), which does not need to generate candidate itemsets and is considered more efficient. FP-growth is constructed by reading the data set one transaction at a time and mapping each transaction onto a path in a Frequent Pattern-tree (FP-tree). Since different transactions can have several items in common at the same time, their paths may overlap. The more paths overlap with one another, the more compression we can achieve by using the FP-tree structure. If the size of the FP-tree is small enough to fit into the main memory, we can extract frequent itemsets directly from the structure in memory instead of making repeated passes over the data stored on disks [13]. Therefore, without generating candidate itemsets, one only needs to scan the database twice.

Additionally, in regard to the matter of decision making, one has to take users' perception and cognitive uncertainty of subjective decisions into consideration. Zadeh proposed the Fuzzy Set Theory [15] in 1965 to deal with cognitive uncertainty of vagueness and ambiguity. Since linguistic variables and linguistic values [16-18] can be described with fuzzy concepts to subjectively correspond with the possible cognition of a decision maker, they are handy in carrying out analysis of decisions-making. Fuzzy data mining has then recently become an important research matter.

Therefore, this paper proposes a fuzzy data mining method - Fuzzy Frequent Pattern Growth, which treats each item from a transaction database as a linguistic variable, and each linguistic variable is partitioned based on its linguistic value. In so doing, the natural language can be utilized to fully explain fuzzy association rules. There are two phases in the proposed method. One is to find frequent 1-itemset by scanning the database once, and the other is to establish a membership function FP-tree by scanning the database twice. Then, one conditional pattern base and one conditional membership function FP-tree will be extracted from each node in a membership function FP-tree to generate the fuzzy association rules.

## II. FUZZY PARTITION METHOD

The concept of linguistic variables plays a fundamental role in describing fuzzy logic and approximate reasoning. To deal with linguistic variables, the fuzzy partition method [9-11, 14] can divide quantitative variables into fuzzy sets with membership functions. For example, a linguistic "age" can be partitioned into young, middle and old. And their membership functions in [0, 60] are (0, 0, 30), (0, 30, 60), (30, 60, 60), respectively.

Common membership functions are triangular or trapezoid membership functions. In this paper, symmetric triangle-shaped linguistic variables are used for simplicity. We assume that a quantitative variable a is partitioned into $K$ fuzzy sets $\{A_{k,i_1}, A_{k,i_2}, \ldots, A_{k,i_k}\}$, $A_{k,i_k}$ is the $i$th fuzzy grid and is defined by the triangular membership function [7-10]:

$$\mu_{k,i_k}(x) = \max\{1 - |x - a_{i_k}^{K_i}| / b^{K_i}, 0\}, \quad (1)$$

where $a_{i_k}^{K_i} = m_i + (m_a - m_i)(i_k - 1)/(K_k - 1)$, $b^{K_i} = (m_a - m_i)/(K_k - 1)$, and $m_a$ and $m_i$ are the maximum and minimum values of the domain interval of $A_{k,i_k}$. $a_{i_k}^{K_i}$ is the top where the membership degree is equal to 1 and $b^{K_i}$ is the spread of the membership function of $A_{k,i_k}$.

In the data set we use fuzzy partition method to transform quantitative attributes into fuzzy grids. A quantitative attribute $x_k$ is represented as:

$$\left(\frac{\mu_{k,i_1}}{A_{k,i_1}} + \frac{\mu_{k,i_2}}{A_{k,i_2}} + \cdots + \frac{\mu_{k,i_k}}{A_{k,i_k}}\right) \quad (2)$$

using the triangular membership function defined by above formula, where $A_{k,i_k}$ is the $i$th fuzzy grid of $K$ linguistic terms defined in attribute $x_k$ of $p$th transaction data.

## III. NOTATIONS AND ALGORITHM

The proposed construction algorithm for generating fuzzy association rules from the transaction database is described in this section. The notations used in the proposed algorithm are firstly stated below. An example, e.g. TABLE I is given to illustrate the proposed algorithm behind steps. The assumptions of the membership functions for the item quantities are shown in Fig. 1. Additionally, assume that the predefined minimum fuzzy support value and minimum fuzzy confidence value are 0.24 and 0.55, respectively.

### A. Notations

- $n$: rhe number of transaction database;
- $d$: mumber of attributes used to describe each sample data, where $1 \le m$;
- $x_i$: $i$th attribute;
- $K_i$: mumber of linguistic values in the $i$th attribute where $1 \le i \le d$;
- $A_{k,i_k}$: $i_k$th linguistic value of $K_k$ linguistic values defined in linguistic variable $x_k$, where $1 \le i \le K$;
- $\mu_{k,i_k}$: membership function of $A_{k,i_k}$;
- $t_p$: $p$th transaction data;
- $count_{k,i_k}$: the summation of $\mu_{k,i_k}$ value for $i$ =1 to $n$;
- $count_k^{\max}$: the maximum count value among $count_{k,i_k}$ value, $i_k$ = 1 to $K$;
- $G_k^{\max}$: the fuzzy grid $x_i$ with $count_k^{\max}$;
- $\alpha$: the user-specified minimum fuzzy support value;
- $\beta$: the user-specified minimum fuzzy confidence value;
- $SP_r$: the set of frequent lengths with $r$ attributes.

### B. The proposed algorithm

INPUT: A body of $n$ transaction data; each linguistic variable with $K$ linguistic values; a user-specified minimum fuzzy support (min FS) $\alpha$; and a user-specified minimum fuzzy confidence (min FC) $\beta$.

OUTPUT: A set of fuzzy association rules.

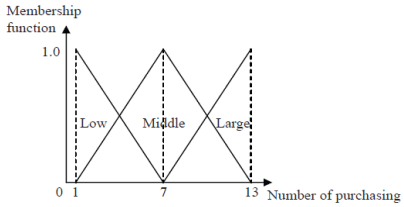


Fig. 1. The membership functions are used in the example.

TABLE I. THE DATA SETS ARE USED IN THIS EXAMPLE

| TID | ITMS |
|---|---|
| 1 | (A, 3), (B, 8), (D, 4), (E, 11) |
| 2 | (B, 7), (C, 2), (D, 3) |
| 3 | (A, 2), (D, 7), (E, 6) |
| 4 | (A, 3), (B, 2), (E, 5) |
| 5 | (B, 5), (C, 3), (D, 8) |

STEP 1: Transform the quantitative attribute $A_{k,i_k}$ for each item $x_k$ in each transaction datum $t_p$ ($p$ =1 to $n$) into a fuzzy set $\mu_{k,i_k}$. For example, the results are shown in TABLE II.

STEP 2: Calculate the count of each fuzzy grid (linguistic term) $\mu_{k,i_k}$ in the transaction data, and the results are shown in TABLE III.

$$count_{k,i_k} = \sum_{p=1}^{m} \mu_{k,i_k} \quad (3)$$

TABLE II. THE FUZZY SETS TRANSFORMED FROM THE DATA SET IN TABLE 1.

| TID | ITMS |
|---|---|
| 1 | (0.667/A.Low + 0.333/A.Middle), (0.833/B.Middle + 0.167/B.Large) + (0.5/D.Low + 0.5/D.Middle) + (0.333/E.Middle + 0.667/E.High) |
| 2 | (1.0/B.Middle), (0.833/C.Low + 0.167/C.Middle), (0.667/D.Low + 0.333/D.Middle) |
| 3 | (0.833/A.Low + 0.167/A.Middle), (1.0/D.Middle), (0.667/E.Low + 0.333/E.Middle) |
| 4 | (0.667/A.Low + 0.333/A.Middle), (0.833/B.Low + 0.167/B.Middle), (0.333/E.Low + 0.667/E.Middle) |
| 5 | (0.333/B.Low + 0.667/B.Middle), (0.667/C.Low + 0.333/C.Middle), (0.833/D.Middle + 0.167/D.Large) |

TABLE III. THE COUNT OF THE FUZZY GRIDS.

| Item | Count | Item | Count |
|---|---|---|---|
| A.Low | 2.167 | C.Large | 0 |
| A.Middle | 0.833 | D.Low | 1.167 |
| A.Large | 0 | D.Middle | 2.667 |
| B.Low | 1.167 | D.Large | 0.167 |
| B.Middle | 2.667 | E.Low | 0.5 |
| B.Large | 0.167 | E.Middle | 1.833 |
| C.Low | 1.5 | E.Large | 0.667 |
| C.Middle | 0.5 | | |

STEP 3: Find $count_k^{\max}$ . Let $G_k^{\max}$ be the grid with $count_k^{\max}$ for item $x_i$. $G_k^{\max}$ will be used to represent this attribute in later FP-Growth mining processing. This step is repeated for the other items. Thus, "*Low*" is chosen for *A*, "*Middle*" is chose for *B*, "*Low*" is chosen for *C*, "*Middle*" is chosen for *D* and "*Middle*" is chosen for *E*.

STEP 4: Collect each attribute grid to form the 1-dim pattern ($SP_1$).

STEP 5: Check whether each $count_k^{\max}$ of each $G_k^{\max}$ is larger than or equal to the predefined minimum support value α. If $G_k^{\max}$ satisfies the above condition, put it in the set of frequent lengths. That is:

$$SP_1 = \{G_k^{\max} \mid count_k^{\max} \geq \alpha,\ 1 \leq i \leq K\} \qquad (4)$$

Here, minimum fuzzy support is 0.24, then its count is 1.2. So the counts values of *A.Low*, *B.Middle*, *C.Low*, *D.Middle* and *E.Middle* are all larger than 1.2, these 1-dim patterns are put in $SP_1$.

STEP 6: While frequent fuzzy grids ($SP_1$) are not null, do the following steps.

STEP 7: This step begins to proceed with FP-Growth. Establish a descending data table called Header Table, shown as TABLE IV.

TABLE IV. HEADER TABLE.

| 1-dim pattern | Count |
|---|---|
| B.Middle | 2.667 |
| D.Middle | 2.667 |
| A.Low | 2.167 |
| E.Middle | 1.833 |
| C.Low | 1.5 |

STEP 8: Scan the fuzzy set table at the first time to rebuild a new fuzzy set table, e.g. TABLE V, that the fuzzy sets ate sorted by the fuzzy grids from the Header Table.

STEP 9: Set a node {ROOT} of a fuzzy FP-tree. Then the second time to scan the new fuzzy set table could generate the nodes of the fuzzy grids in the tree then link the nodes with one another based on the same transactions. In the example, the fuzzy FP-tree is shown in Fig. 2.

TABLE V. THE NEW FUZZY SETS FROM TABLE 1.

| TID | Fuzzy sets |
|---|---|
| 1 | (0.833/B.Middle), (0.5/D.Low), (0.667/A.Low), (0.333/E.Middle) |
| 2 | (1.0/B.Middle), (0.333/D.Low), (0.833/A.Low) |
| 3 | (1.0/D.Middle), (0.833/A.Low), (0.333/E.Middle) |
| 4 | (0.167/B.Middle), (0.667/A.Low), (0.667/E.Middle) |
| 5 | (0.667/B.Middle), (0.833/D. Middle), (0.333/C.Low) |

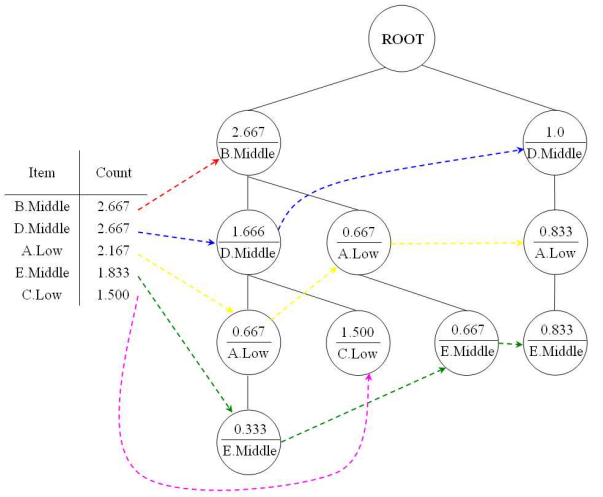


Fig. 2. The membership functions for FP-tree.

STEP 10: Mine the possible frequent patterns through fuzzy FP-tree. Scanning from the bottom of Header Table derives conditional pattern bases of every item from the paths in the fuzzy FP-tree. Next, the conditional membership function FP-trees of every item could be constructed based on the conditional pattern bases. Then, the whole set of possible patterns of each item would be generated from the corresponding conditional fuzzy FP-tree.

STEP 11: The following substeps are done for corresponding frequent patterns.

(a) Calculate the fuzzy value of each transaction data tp in s as $\mu_s^i = \mu_{s_1}^i \Lambda \mu_{s_2}^i \ldots \Lambda \mu_{s_{r+1}}^i$ , where $\mu_s^i$ is the membership function of $t_p$ in grids $s_k$. and the minimum operator is used for the intersection , then

$$\mu_s^i = \min_{k=1}^{r+1} \mu_{s_k}^i \qquad (4)$$

(b) If fuzzy support for each frequent pattern is calculated as

$$count_s = \sum_{i=1}^{n} \mu_s^i \quad (4)$$

(c) If fuzzy support value is larger or equal to the user-specified minimum support value α, put it is $SP_r$.

In the example, the possible patterns of (*B.Middle, C.Low*), (*A.Low, E.Middle*), (*D.Middle, A.Low*), (*B.Middle, D.Middle*) are kept in $SP_2$. Therefore, they are frequent patterns.

STEP 12: Construct effective association rules for each frequent pattern ( $s_1, s_2, ..., s_q, q \geq 2$ ) using the following.

(a) List all possible frequent patterns.

$$s_1 \Lambda s_2 \ldots \Lambda s_q \rightarrow s_l, l = 1 \text{ to } q$$

(b) Calculate the confidence values of all association rules using

$$\frac{\sum_{i=1}^{n} \mu_s^i}{\sum_{i=1}^{n} (\mu_{s_1}^i \Lambda \mu_{s_2}^i \Lambda \cdots \Lambda \mu_{s_q}^i)} \quad (5)$$

STEP 13: Output the association rules with confidence values larger than or equal to the user-specified confidence threshold $\beta$. Here, the min fuzzy confidence is 0.55. Since (*A.Low, E.Middle*) is larger than 0.55, therefore this frequent pattern is an effective rule. In fact, the following four frequent patterns form association rules and are an output to users.

1. (*C.Low, B.Middle*) with fuzzy confidence = 1.0: If a **low** number of *C* is bought then a **middle** of *B* is bought with a confidence of 1.0.

2. (*A.Low, E.Middle*) with fuzzy confidence = 0.846: If a **low** number of *A* is bought then a **middle** of *E* is bought with a confidence of 0.846.

3. (*E.Middle, A.Low*) with fuzzy confidence = 1.0: If a **middle** number of *E* is bought then a **low** of *A* is bought with a confidence of 1.0.

4. (*A.Low, D.Middle*) with fuzzy confidence = 0.615: If a **low** number of *A* is bought then a **middle** of *D* is bought with a confidence of 0.615.

The four rules above are the output as meta-knowledge concerning the given transaction.

## IV. Experimental Results

The section reports on experiments made to show the performance of the proposed approach. They were implemented in VB on a Pentium-IV 1.7 Personal Computer. Transaction database from a supermarket of a retail business in Kinmen, Taiwan, were used to show the feasibility. A total of 4862 transactions were included in the database. Each transaction records the purchasing information for a customer and uses membership function as Fig. 3.

The experimental results were made to compare the accuracy of the proposed fuzzy data mining algorithm. Hong et al.'s and Hu's proposed algorithm ([5, 6]) in which the minimum fuzzy support values set at 0.3 is shown in Figure 3. Also, we simulate even one hundred thousand to compare these three algorithms which set minimum fuzzy confidence at 0. The results are shown in TABLE VI.

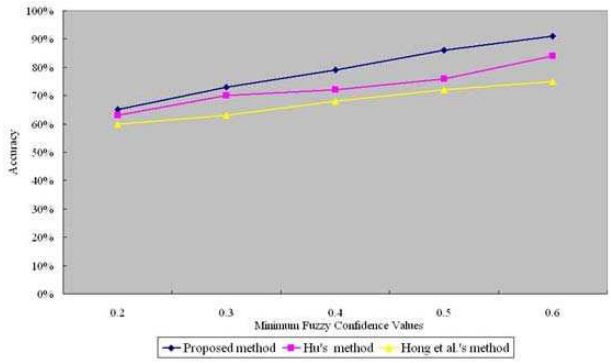


Fig. 3. The comparison of the accuracy of three fuzzy data mining algorithms.

TABLE VI. The Experiment Results for Three Data Minings Algorithms.

| Min FS | The proposed method | | Hu's mehod | | Hong et al's method | |
|---|---|---|---|---|---|---|
| | Time(s) | Rule(s) | Time(s) | Rule(s) | Time(s) | Rule(s) |
| 0.2 | 965 | 1327 | 1358 | 1286 | 1873 | 1083 |
| 0.25 | 821 | 962 | 1027 | 997 | 1562 | 852 |
| 0.3 | 625 | 534 | 753 | 624 | 1194 | 708 |
| 0.35 | 292 | 186 | 571 | 268 | 856 | 427 |
| 0.4 | 123 | 45 | 264 | 63 | 577 | 199 |
| 0.45 | 28 | 8 | 86 | 15 | 312 | 22 |

From Fig. 3, it is easily seen that the accuracy of the proposed method was higher than that of Hong et al and Hu's proposed method for various minimum fuzzy confidence values. In addition, from TABLE VI, it is easily seen that minimum fuzzy support is smaller, which generates more association rules. In execute time, the proposed method is the best than the other methods, the minimum fuzzy support is smaller, especially. But smaller minimum fuzzy support does not influence the whole execute time. With the minimum fuzzy support increases, the execute time will be more obvious.

## V. Conclusions

In this paper, we have proposed a fuzzy data mining algorithm, which combines fuzzy set theory and FP-Growth to deal with quantitative values and find interesting patterns among them. The rules mined represent quantitative regularity for large transaction database and can be adopted to provide some suggestions to appropriate supervisors. The proposed algorithm can solve disadvantages found in Apriori and enhance the whole efficiency. The experimental results with the data in a supermarket of retail business show the feasibility of the proposed mining algorithm. When compared with Hong et al. and Hu's fuzzy mining method ([5, 6]), our approach can get better mining results due to the characteristic of FP-growth. It owns feature of without candidate generation, and scans two for all process. The experimental results pronounced the proposed method is more excellent than Apriori mining method.

In addition, as for definitions of linguistic values, managers can select his own preference, and refer to past experiences and relate cognitive ability to design a number of linguistic values and shapes, such as Gaussian distribution and trapezoid membership function. So, it confirms a manager's cognitive in subject. However, it is not easy for users to specify these parameters, the GA may be considered as an appropriate tool for automatically determining the appropriate parameter specifications. As regard to min FS, min FC and partition numbers in each quantitative attribute, these parameters are easily user-specified for individual problems. And GA owns to automatically search for characteristic. In the future, we will combine GA and the proposed method to solve various problem domains.

## *References*

[1] Berry, M., Linoff, G., Data Mining Techniques: for Marketing, Sales, and Customer Support, John Wiley & Sns, Now York, 1997.

[2] Han, J. W., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[3] Han, J., Pei, J., Yin, Y., "Mining frequent patterns without candidate generation," In proceeding SIGMOD'00 Proceeding of the 2000 ACM SIGMOD international conference on Management of data, pp. 1-12, 2000.

[4] Hong, T. P., Chen, J. B., "Find relevant attributes and membership functions," Fuzzy Sets and Systems, vol. 103, pp.389-404, 1999.

[5] Hong, T. P., Kuo, C. S., Chi, S. C., "Trade-off between computation time and number of rules for fuzzy mining from quantitative data," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 9, pp. 587-604, 2001.

[6] Hu, Y. C., "Mining association rules at a concept hierarchy using fuzzy partition," Journal of Information Management, vol. 13, p. 63-80, 2006.

[7] Hu, Y. C., Chen, R. S., and Tzeng, G. H., "Finding fuzzy classification rules using data mining techniques," Pattern Recognition Letters, vol. 24, pp. 509-519, 2003.

[8] Hu, Y. C., "Finding useful fuzzy concepts for pattern classification using genetic algorithm," Information Sciences, vol. 175, pp. 1-19, 2005.

[9] Ishibuchi, H., Nozaki, K., Yamaoto, N., "Tanaks, Selecting fuzzy if-then rules for classification problem using genetic algorithm," IEEE Transactions on Fuzzy Systems, vol. 3, pp. 260-270, 1995.

[10] Ishibchi, H., Nakashima, T., Murata, T., "Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems," IEEE Transaction on Systems, Man, and Cybernetics, vol. 29, pp. 601-618, 1999.

[11] Jang, J.S.R., "ANFIS: adaptive-network-based fuzzy inference systems," IEEE Transactions on Systems, Man, and Cybernetics, 23, pp. 665-685, 1993.

[12] Myra, S., "Web usage mining for web site evaluation," Communication of the ACM, 43, pp. 127-134, ACM, New York, 2000.

[13] Tan, P. N. , Steinbach, Michael, Kumar Vipin, Introduction to Data Mining, Pearson Addison Wesley, Boston, 2005.

[14] Wang, L. X., Mendel, J. M., "Generating fuzzy rules by learning from examples. IEEE Transactions on Systems," Man, and Cybernetics, vol. 22, pp. 1414-1427, 1992.

[15] Zadeh, L. A., "Fuzzy Sets," Information Control, vol. 8, pp. 338-353, 1965.

[16] Zadeh, L. A., "The concept of a linguistic variable and its application t approximate reasoning – I," Information Science, vol. 8, pp. 199-249, 1975[a].

[17] Zadeh, L. A., "The concept of a linguistic variable and its application t approximate reasoning – II." Information Science, vol. 8, pp. 301-357, 1975[b].

[18] Zadeh, L. A., "The concept of a linguistic variable and its application t approximate reasoning – III," Information Science, vol. 9, pp. 43-80, 1976.