Fuzzy Correlation Rules Mining

NANCY P. LIN, HAO-EN CHUEH
Department of Computer Science and Information Engineering,
Tamkang University,
151 Ying-chuan Road, Tamsui, Taipei,
TAIWAN

Abstract: - General fuzzy association rules mining focuses on finding out the fuzzy itemsets or fuzzy attributes which frequently occur together. But two fuzzy itemsets which frequently occur together can not imply that there is always an interesting relationship between them. In this paper, we develop an alternative framework for mining interesting relationship between fuzzy itemsets based on fuzzy correlation analysis, and the discovered rules are called fuzzy correlation rules. The analysis of fuzzy correlation can show us the strength and the type of the linear relationship between two fuzzy itemsets, and hence can prevent generating the misleading rules.

Key-Words: - Fuzzy association rules, Fuzzy itemsets, Fuzzy correlation rules, Fuzzy correlation analysis, Linear relationship, Misleading rules

1 Introduction

Data mining is often defined as finding hidden information from a large dataset [14, 17], and one of the popular data mining tasks is association rule mining [1, 2, 3, 21, 22]. Association rules mining is a process to find out the itemsets which frequently occur together from a large dataset. It is widely used in retail business to identify products that are frequently purchased together. Clearly, this kind of information is useful for assisting in making marketing decisions.

However, in practical datasets, data may be fuzzy but useful, waiting to be explored, and methods to find out association rules from these fuzzy datasets are certainly needed. To reach this goal, the fuzzy sets theory is commonly used, and the discovered rules are called fuzzy association rules [6, 9, 12].

Many fuzzy association rules mining algorithms have been proposed for various fuzzy dataset [5, 15, 16, 19, 20, 24], and most algorithms employ a support-confidence framework. In these algorithms, minimal support and minimal confidence are used to choose interesting fuzzy association rules from the fuzzy itemsets which frequently occur together.

However, a situation needs to be thought with here, if a fuzzy itemset almost occurs in all records, then it most likely occurs with other fuzzy itemsets frequently, but the relationship between these fuzzy itemsets may be not interesting in fact. Therefore, an alternative framework to prevent generating misleading fuzzy rules is a must. Hence, in this paper, we develop a novel algorithm for mining interesting relationship between fuzzy itemsets based on fuzzy correlation analysis [8, 10, 18, 23], and thus, the discovered fuzzy rules are called fuzzy correlation rules.

This paper is organized as follows: In section 2, some basic concepts of the fuzzy association rules mining are mentioned. In section 3, the main concept of our proposed fuzzy correlation rules mining algorithm, the fuzzy correlation analysis, is introduced. In section 4, how to use the fuzzy correlation analysis in our algorithm is explained. In section 5, an example of fuzzy correlation rules mining is given. Section 6 is our conclusion.

2 Fuzzy Association Rules Mining

Fuzzy association rules mining is a process to find out the fuzzy itemsets or fuzzy attributes which frequently occur together from a fuzzy dataset [5, 6, 9, 12, 15, 16, 19, 20, 24].

Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of fuzzy items, $T = \{t_1, t_2, \cdots, t_n)$ be a set of fuzzy data records, and each record t_i is represented as a vector with m values, $(f_1(t_i), f_2(t_i), \cdots, f_m(t_i))$, where $f_j(t_i)$ is membership degree that t_i belongs to fuzzy item f_j, and $f_j(t_i) \in [0,1]$. A fuzzy association rule is defined as an implication of the form $F_X \rightarrow F_Y$, where $F_X, F_Y \subset F$ are fuzzy itemsets, $\forall_{x,y} f_x \in F_X \neq$

$f_y \in F_Y$.

A fuzzy association rule, say $F_X \rightarrow F_Y$, holds in fuzzy dataset T with fuzzy support ($fsupp(\{F_X, F_Y\})$) and with fuzzy confidence ($fconf(F_X \rightarrow F_Y)$) defined as follows:

$$fsupp(\{F_X, F_Y\}) = \frac{\sum_{i=1}^{n} \min(f_j(t_i) \mid f_j \in \{F_X, F_Y\})}{n} \quad (1)$$

$$fconf(F_X \rightarrow F_Y) = \frac{fsupp(\{F_X, F_Y\})}{fsupp(F_X)} \quad (2)$$

If $fsupp(\{F_X, F_Y\})$ is greater than or equal to the user-predefined minimal fuzzy support (s_f) and $fconf(F_X \rightarrow F_Y)$ is also greater than or equal to the user-predefined minimal fuzzy confidence (c_f), then, fuzzy association rule $F_X \rightarrow F_Y$ is considered as an interesting fuzzy association rule, and it means that F_X and F_Y frequently occur together in same records.

Now, let us consider a special case, if a fuzzy itemset is common, and it almost occurs in all fuzzy data records, then according to the formula (1) and formula (2), we most likely obtain some interesting fuzzy association rules concerning this fuzzy itemset. But, in fact, the occurrence of this fuzzy itemset does not imply the occurrence of other fuzzy itemset, the discovered rules are misleading.

Some researchers have noticed this problem, and thus turned to adopt alternative measures which can show extra information about the relationships between itemsets in mining processes [7, 14, 11, 13, 17, 22]. A most famous measure is stated as follows [7, 14, 17].

Support there are two itemsets, A and B, and the probability that A occurs in the given dataset is expressed as $P(A)$; the probability that B occurs is expressed as $P(B)$; the probability that A and B that occur together is expressed as $P(A,B)$. Then the correlation of the association rule $A \rightarrow B$ can be expressed as *correl*($A \rightarrow B$).

$$correl(A \rightarrow B) = \frac{P(A,B)}{P(A) \cdot P(B)} \quad (3)$$

The value derived from formula (3) lies between in $[0, \infty]$. If *correl*($A \rightarrow B$) is greater than 1, than A and B are positively related, meaning the occurrence of one implies the occurrence of the other; if *correl*($A \rightarrow B$) is less than 1, than the occurrence of one is negatively related with the occurrence of the other; the occurrence of one is independent of the occurrence of the other if $P(A,B) = P(A) \cdot P(B)$.

However, this kind of probability-based formula is not suitable to analyze the relationship between fuzzy itemsets, and thus, in order to find out the interesting relationships between fuzzy itemsets, in this paper, the fuzzy correlation analysis defined in section 3 [10] is adopted to construct a novel fuzzy correlation rules mining algorithm. The concepts of the fuzzy correlation analysis and how to use the fuzzy correlation analysis in our proposed algorithm will be next explained.

3 Fuzzy Correlation Analysis

The coefficient of correlation analysis of fuzzy sets is called fuzzy correlation coefficient. Many methods have been proposed to calculate the fuzzy correlation coefficient [8, 10, 18, 23]; here, we adopt the formula derived by Lin [10], because it can provide the extra information we need.

Suppose there are two fuzzy itemsets $A, B \subset F$, where F is a fuzzy space. A and B are defined on a crisp universal set X with membership functions μ_A and μ_B, and thus the fuzzy itemsets A and B can be expressed as follows:

$$A = (x, \mu_A(x)) \mid x \in X) \quad (4)$$

$$B = (x, \mu_B(x)) \mid x \in X) \quad (5)$$

where $\mu_A, \mu_B \in [0, 1]$.

Assume that there is a random sample (x_1, $x_2, \cdots, x_n$) $\in X$, alone with a sequence of paired data, $\{(x_i, \mu_A(x_i), \mu_B(x_i)) \mid i = 1 \cdots n\}$, which correspond to the grades of the membership functions of fuzzy itemsets A and B defined on X. Then the fuzzy correlation coefficient between the fuzzy itemsets A and B, $r_{A,B}$ is:

$$r_{A,B} = s_{A,B} \Big/ \sqrt{s_A^2 \cdot s_B^2} \quad (6)$$

where

$$s_{A,B} = \frac{\sum_{i=1}^{n}(\mu_A(x_i) - \overline{\mu_A}) \cdot (\mu_B(x_i) - \overline{\mu_B})}{n-1} \quad (7)$$

$$\overline{\mu_A} = \frac{\sum_{i=1}^{n} \mu_A(x_i)}{n} \quad (8)$$

$$\overline{\mu_B} = \frac{\sum_{i=1}^{n} \mu_B(x_i)}{n} \quad (9)$$

$$S_A{}^2 = \frac{\sum_{i=1}^{n} (\mu_A(x_i) - \overline{\mu_A})^2}{n-1} \quad (10)$$

$$S_B{}^2 = \frac{\sum_{i=1}^{n} (\mu_B(x_i) - \overline{\mu_B})^2}{n-1} \quad (11)$$

$$S_A = \sqrt{S_A{}^2} \quad (12)$$

$$S_B = \sqrt{S_B{}^2} \quad (13)$$

The value derived from (6) lies between in [-1, 1], and some important properties of fuzzy correlation coefficient are stated as follows [4, 10]:

1. If $|r_{A,B}|$ is close to 1, then the fuzzy itemsets A and B are highly related.
2. If $|r_{A,B}|$ is close to 0, then the fuzzy itemsets A and B are barely related.
3. If $r_{A,B} > 0$, then the fuzzy itemsets A and B are positively related.
4. If $r_{A,B} < 0$, then the fuzzy itemsets A and B are negatively related.
5. If $r_{A,B} = 0$, then the fuzzy itemsets A and B have no relationship at all.

According to the above properties, fuzzy correlation coefficient is great useful for mining the interesting relationship between fuzzy itemsets. Next, how to use the fuzzy correlation analysis in our fuzzy correlation rules mining algorithm will be accounted for.

4 Fuzzy Correlation Rules Mining

In this section, our proposed fuzzy correlation rules mining algorithm is explained.

Assume that $F = \{f_1, f_2, \cdots, f_m\}$ be a set of fuzzy items; $T = \{t_1, t_2, \cdots, t_n)$ be a set of fuzzy data records, and each record t_i is represented as a vector with m values, $(f_1(t_i), f_2(t_i), \cdots, f_m(t_i))$, where $f_j(t_i)$ is membership degree that t_i belongs to fuzzy item f_j (i.e., $f_j(t_i) = \mu_{f_j}(t_i)$), $f_j(t_i) \in [0,1]$; s_f is the user-predefined minimal fuzzy support; c_f is the user-predefined minimal fuzzy confidence; α is the user-predefined minimal fuzzy correlation. Then, the process of our proposed fuzzy correlation rules mining algorithm is described as the following steps:

Step 1: For each fuzzy item $f_i \in F$, $fsupp(f_i)$ is computed.

Step 2: Let $L_1 = \{F_i \mid F_i \in F,\ fsupp(F_i) \geq s_f\}$ is the set of frequent fuzzy itemsets whose size is equal to 1.

Step 3: Let $C_2 = \{(F_A, F_B)\}$ is the set of candidate combinations of two fuzzy itemsets of L_1, where $F_A, F_B \in L_1$ and $F_A \neq F_B$, that is, C_2 is generated by L_1 joint with L_1. Because F_A and F_B are the elements of L_1, the number of the fuzzy items of each element of C_2 is 2.

Step 4: For each element of C_2, say (F_A, F_B), the fuzzy support (*fsupp* $(\{F_A, F_B\})$) and the fuzzy correlation coefficient between F_A and F_B $(r_{A,B})$ are computed. If *fsupp* $(\{F_A, F_B\})$ is greater than or equal to s_f , and $r_{A,B}$ is greater than or equal to α, then the combination (F_A, F_B) is an element of L_2. Hence, L_2 is the set of large (or frequent) combinations of two fuzzy itemsets of L_1.

Step 5: Next, each $C_k, k \geq 3$, can be generated by L_{k-1} joint with itself. Suppose (F_W, F_X) and $(F_Y,\ F_Z)$ are two elements of L_{k-1}, and one of the fuzzy itemsets of (F_W, F_X), say F_X , is equal to one of the fuzzy itemsets of $(F_Y,\ F_Z)$, say F_Y, and the total number of the fuzzy items of the combination $(F_X,\ \{F_W, F_Z\})$ is equal to k , and $(F_W,\ F_Z)$ is also a large combination of two fuzzy itemsets, then the combination $(F_X, \{F_W, F_Z\})$ will be a element of C_k. Next, for each element of C_k, the fuzzy support and the fuzzy correlation coefficient are still used to select the elements of L_k.

Step 6: When each L_k , $k \geq 2$, is obtained, for each element of L_k, say (F_I, F_J), two fuzzy rules, $F_I \rightarrow F_J$ and $F_J \rightarrow F_I$, can be generated. If the fuzzy confidence of a rule is greater than or equal to c_f , then it is considered as an interesting fuzzy correlation rule.

The algorithm won't stop until no next C_{k+1} can be generated. A simple example is displayed in the next section.

5 Example

A sample fuzzy dataset is shown as in Table 1. $F=\{f_1, f_2, f_3, f_4, f_5\}$, and $T=\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}\}$. Assume that s_f is set to 0.30, c_f is set to 0.75, and α is set to 0.4.

Table 1: A sample fuzzy dataset.

T \ F	f_1	f_2	f_3	f_4	f_5
t_1	0.1	0.9	0.3	0.5	0.3
t_2	0.2	0.8	0.4	0.8	0.4
t_3	0.2	0.7	0.1	0.3	0.5
t_4	0.4	0.5	0.6	0.1	0.1
t_5	0.7	0.3	0.4	0.8	0.2
t_6	0.5	0.6	0.7	0.9	0.8
t_7	0.8	0.1	0.9	0.4	0.9
t_8	0.7	0.1	0.5	0.2	0.7
t_9	0.8	0.4	0.3	0.5	0.6
t_{10}	0.2	0.7	0.2	0.4	0.5

First, the fuzzy support of each fuzzy item of F is computed and listed in Table 2.

Table 2: The fuzzy support of each fuzzy item of F.

F	fuzzy support
f_1	0.46
f_2	0.51
f_3	0.44
f_4	0.49
f_5	0.50

Because $fsupp(f_i), i=1\cdots5$, are all greater than s_f, we obtain the set of the frequent fuzzy itemsets whose size is equal to 1, L_1. $L_1=\{f_1, f_2, f_3, f_4, f_5\}$.

Next, the set of candidate combinations of two fuzzy itemsets of L_1, C_2, is generated by L_1 joint with itself. Thus, $C_2=\{(f_1, f_2), (f_1, f_3), (f_1, f_4), (f_1, f_5), (f_2, f_3), (f_2, f_4), (f_2, f_5), (f_3, f_4), (f_3, f_5), (f_1, f_4)\}$.

For each element of C_2, the fuzzy support and the fuzzy correlation coefficient are computed and listed in Table 3.

Table 3: The fuzzy support and the fuzzy correlation coefficient of each element of C_2.

C_2	fuzzy support	fuzzy correlation coefficient
$(\{f_1\},\{f_2\})$	0.25	-0.91
$(\{f_1\},\{f_3\})$	0.35	0.54
$(\{f_1\},\{f_4\})$	0.31	0.01
$(\{f_1\},\{f_5\})$	0.36	0.44
$(\{f_2\},\{f_3\})$	0.29	-0.56
$(\{f_2\},\{f_4\})$	0.36	0.25
$(\{f_2\},\{f_5\})$	0.32	-0.41
$(\{f_3\},\{f_4\})$	0.31	0.08
$(\{f_3\},\{f_5\})$	0.37	0.43
$(\{f_4\},\{f_5\})$	0.36	0.10

In Table 3, an element whose fuzzy support is greater than or equal to s_f and fuzzy correlation coefficient is greater than or equal to α is considered an element of L_2.

$L_2=\{(\{f_1\},\{f_3\}),(\{f_1\},\{f_5\}),(\{f_3\},\{f_5\})\}$.

When L_2 is obtained, C_3 can be generated by L_2 joint with L_2. $C_3=\{(\{f_1\},\{f_3, f_5\}), (\{f_3\},\{f_1, f_5\}), (\{f_5\}, \{f_1, f_3\})\}$.

Similarly, the fuzzy support and the fuzzy correlation coefficient of each element of C_3 are also computed and displayed in Table 4.

Table 4: The fuzzy support and the fuzzy correlation coefficient of each element of C_3.

C_3	fuzzy support	fuzzy correlation coefficient
$(\{f_1\},\{f_3, f_5\})$	0.30	0.47
$(\{f_3\},\{f_1, f_5\})$	0.30	0.57
$(\{f_5\},\{f_1, f_3\})$	0.30	0.55

In Table 4, because all elements of C_3 satisfy the minimal fuzzy support (s_f) and the minimal fuzzy correlation (α) , all elements of C_3 are elements of L_3 .

L_3 ={({ f_1 },{ f_3 , f_5 }), ({ f_3 },{ f_1 , f_5 }), ({ f_5 }, { f_1 , f_3 })}.

No next C_4 can be generated by L_3 joint with L_3 , so our algorithm stops here. Only 2 sets of frequent combinations of two fuzzy itemsets are obtained, L_2 and L_3 , From L_2 and L_3 , 14 candidate fuzzy correlation rules can be generated. The fuzzy confidences of these rules are shown as in Table 5.

Table 5: The fuzzy confidences of the candidate fuzzy correlation rules.

C_2	fuzzy confidence
{ f_1 }→{ f_3 }	0.76
{ f_3 }→{ f_1 }	0.80
{ f_1 }→{ f_5 }	0.78
{ f_5 }→{ f_1 }	0.72
{ f_2 }→{ f_4 }	0.71
{ f_4 }→{ f_2 }	0.74
{ f_3 }→{ f_5 }	0.84
{ f_5 }→{ f_3 }	0.74
{ f_1 }→{ f_3 , f_5 }	0.65
{ f_3 , f_5 }→{ f_1 }	0.81
{ f_3 }→{ f_1 , f_5 }	0.68
{ f_1 , f_5 }→{ f_3 }	0.83
{ f_5 }→{ f_1 , f_3 }	0.60
{ f_1 , f_3 }→{ f_5 }	0.86

According to Table 5, we determine 7 interesting fuzzy correlation rules as follows, because their fuzzy confidences are greater than or equal to the predefined minimal fuzzy confidence c_f , c_f is 0.75 here.

{ f_1 }→{ f_3 } (14)

{ f_3 }→{ f_1 } (15)

{ f_1 }→{ f_5 } (16)

{ f_3 }→{ f_5 } (17)

{ f_3 , f_5 }→{ f_1 } (18)

{ f_1 , f_5 }→{ f_3 } (19)

{ f_1 , f_3 }→{ f_5 } (20)

From this example, we can clearly see that, the number of frequent combinations of fuzzy itemsets is reduced. For example, the number of the elements of C_2 is 10, and in these elements, the number of elements which satisfy s_f is 8, but after fuzzy correlation coefficient testing, the number of elements which belong to L_2 is 4. Thus, we can conclude that, only really interesting relationships between fuzzy itemsets can be discovered by using our proposed algorithm.

6 Conclusion

In this paper, a new fuzzy correlation rules mining algorithm is proposed. General fuzzy rules mining focus on finding out the fuzzy itemsets or fuzzy attributes which frequently occur together. These discovered rules are called fuzzy association rules. However, two fuzzy itemsets which frequently occur together can not imply that there is always an interesting relationship between them. To deal with this situation, fuzzy correlation analysis is used to assist in discovering the fuzzy correlation rules. According to the example in section 5, we clearly see that, by using our proposed algorithm, fuzzy itemsets which frequently occur together but are not really interesting can be efficiently deleted.

References:

[1] R. Agrawal, T. Imielinski, and A. Swami, Mining Association Rules between Sets of Items in Large Databases, *In the proceeding of the ACM SIGMOD International Conference on Management of Data*, Washington D.C., May 1993, pp.207-216.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast Discovery of Association Rules, *Advances in Knowledge Discovery and Data Mining*, Chapter 12, AAAI/MIT Press, 1995.

[3] R. Agrawal, R. Strikant, Fast algorithms for mining association rules, *In the proceeding of the 20th International Conference on Very*

Large Databases, Santiago, Chile, September 1994, pp.487-499.

[4] S. F. Arnold, *Mathematical Statistics*, Prentice-Hall, New Jersey, 1990.

[5] Wai-Ho Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, *In the proceeding of IEEE World Congress on Computational Intelligence*, 1998, pp.1314 –1319.

[6] P. Bosc, D. Dubois, O. Pivert, and H. Prade, On fuzzy association rules based on fuzzy cardinalities, *In the proceeding of IEEE International Fuzzy Systems Conference*, Melbourne, 2001.

[7] S. Brin, R. Motwani, and C. Silverstein, Beyond market baskets: Generalizing association rules to correlations, *In the proceeding of the ACM SIGMOD International Conference on Management of Data*, 1997, pp. 265-276.

[8] H. Bustince, P. Burillo, Correlation of interval-valued intuitionistic fuzzy sets, *Fuzzy sets and systems*, Vol. 74, 1995, pp.237-244.

[9] K.C.C. Chan, A.K.C. Wong, Mining Fuzzy Association Rules, *In the proceeding of the sixth international conference on Information and knowledge management 1997*, Las Vegas, Nevada, United States, November 1997, pp.209-215.

[10] D. A. Chiang, N. P. Lin, Correlation of Fuzzy Sets, *Fuzzy Sets and Systems*, Vol. 102, 1999, pp.221-226.

[11] M. De Cock, C. Cornelis, and E.E. Kerre, Elicitation of fuzzy association rules from positive and negative examples, *Fuzzy Sets and Systems*, Vol. 149, 2005, pp.73–85.

[12] M. Delgado, N. Marín, D. Sánchez, and M. Vila, Fuzzy Association Rules: General Model and Applications, *IEEE Transactions on Fuzzy Systems*, Vol. 11, No. 2, April 2003, pp.214-225.

[13] D. Dubois, E. Hullermeier and H. Prade, A Note on Quality Measures for Fuzzy Association Rules, *In the proceeding of the 10th International Fuzzy Systems Association World Congress (IFSA-03), Lecture Notes in Artificial Intelligence 2715*, Springer-Verlag, 2003, pp.346-353.

[14] M. H. Dunham, *Data mining, Introductory and Advanced Topics*, Pearson Education, Inc., 2003.

[15] A. Fu, M. Wong, S. Sze, W. Wong, W. Wong, and W. Yu. Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes. *In the proceeding of the First International Symposium on Intelligent Data Engineering and Learning*, Hong Kong, October 1998, pp.263-268.

[16] J.M. de Graaf, W.A. Kosters and J.J.W. Witteman, Interesting Fuzzy Association Rules in Quantitative Databases, *In proceeding of PKDD 2001 (The 5th European Conference on Principles of Data Mining and Knowledge Discovery), Springer Lecture Notes in Computer Science 2168*, Freiburg, Germany, September 2001, pp.140-151.

[17] J. Han, M. Kamber, *Data mining: Concepts and Techniques*, Academic Press, 2001.

[18] D. H. Hong, S. Y. Hwang, Correlation of intuitionistic fuzzy sets in probability spaces, *Fuzzy Sets and Systems*, Vol.75, 1995, pp.77-81.

[19] H. Ishibuchi, T. Nakashima and T. Yamamoto, Fuzzy association rules for handling continuous attributes, *In proceeding of the IEEE International Symposium on Industrial Electronics*, 2001, pp.118-121.

[20] C. M. Kuok, A. Fu, and M. H. Wong, Fuzzy association rules in large databases with quantitative attributes, *In ACM SIGMOD Records*, March 1998.

[21] R. Srikant, R. Agrawal, Mining Generalized Association Rules, *Future Generation Computer Systems*, Volume 13(2-3), December 1997, pp.161-180.

[22] X Wu, C Zhang, S Zhang, Efficient mining of both positive and negative association rules , *ACM Transactions on Information Systems*, Vol. 22 , Issue 3, July 2004, pp.381 – 405.

[23] C. Yu, Correlation of fuzzy numbers, *Fuzzy Sets and Systems*, Vol. 55, 1993, pp.303-307.

[24] W. Zhang, Mining Fuzzy Quantitative Association Rules, *In the proceeding of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 1999, pp.99-102.