African Journal of Mathematics and Computer Science Research Vol. 5(2), pp. 28-33, 15 January, 2012
Available online at http://www.academicjournals.org/AJMCSR
DOI: 10.5897/AJMCSR11.157
ISSN 2006-9731 ©2012 Academic Journals

Full Length Research Paper

A study on a novel method of mining fuzzy association using fuzzy correlation analysis

Karthikeyan T.[1]*, Samuel Chellathurai A.[2] and Praburaj B.[1]

[1]Department of Computer Science, PSG College of Arts and Science, Coimbatore, India.
[2]Department of Computer Science, James College of Engineering and Technology, Nagercoil, India.

Accepted 10 January, 2012

Two different data variables may behave very similarly. Correlation is the problem of determining how much alike the two variables actually are and association rules are used just to show the relationships between data items. Mining fuzzy association rules is the job of finding the fuzzy item-sets which frequently occur together in large fuzzy data set, where the presence of one fuzzy item-set in a record does not necessarily imply the presence of the other one in the same record. In this paper a new method of discovering fuzzy association rules using fuzzy correlation rules is proposed, because the fuzzy support and confidence measures are insufficient at filtering out uninteresting fuzzy correlation rules. To tackle this weakness, a fuzzy correlation measure for fuzzy numbers, is used to augment the fuzzy support-confidence framework for fuzzy association rules. A practical study over the academic behaviour of a particular school is done and some valuable suggestions are given, based on the results obtained.

Key words: Fuzzy association rules, fuzzy item-sets, fuzzy data sets, fuzzy support-confidence, fuzzy correlation measure.

INTRODUCTION

Frequent patterns are patterns that appear in a data set frequently. Finding such frequent patterns (Agrawal et al., 1993) plays an essential role in mining fuzzy associations, fuzzy correlations, and many other interesting relationships among data. Thus, frequent pattern mining has become an important data mining task and a focussed theme in data mining research. Frequent pattern mining searches for recurring relationships in a given data set. The earliest form of frequent pattern mining for association rules is the market basket analysis (Agrawal and Strikant, 1994).

Frequent item set mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision making processes, such as catalogue design, cross marketing, and customer shopping behaviour analysis (Dunham, 2003). But, since most of the real time databases are fuzzy in nature, it is necessary to explore and discover association rules and correlation rules in a fuzzy environment. To this end many researchers have proposed methods for mining fuzzy association rules, from various fuzzy datasets (Au and Chan, 1998; Dubois et al., 2003). In this paper the fuzzy correlation methods proposed by S.T. Liu and C. Kao is used due its advantages in working with fuzzy numbers in a fuzzy environment.

If a fuzzy item set almost occurs in all records, then it may frequently occur with other fuzzy item-sets also (Chan and Wong, 1997). In order to find out useful relationships between the fuzzy item-sets based on fuzzy statistics, fuzzy correlation rules Chiang and Lin (1999) are generated. By using the fuzzy correlation analysis, the fuzzy correlation rules for fuzzy numbers are generated to see that two fuzzy sets not only frequently occur together in same records, but also are related to

*Corresponding author. E-mail: t.karthikeyan.gasc@gmail.com.

each other.

The rest of the paper is organized as follows: Firstly, the concepts of frequent fuzzy item-sets, closed fuzzy item-sets and fuzzy association rules, the concept of mining fuzzy association rules, the fuzzy correlation techniques, the mining fuzzy association rules and fuzzy correlation rules are explained. Next, a practical study on the academic behaviour of students (higher secondary level) in a particular subject of a school in Nagercoil, Tamilnadu (India) with the help of the aforementioned discussed methods is presented with careful discussions. This is then followed by conclusions and suggestions.

FREQUENT FUZZY ITEM-SETS, CLOSED FUZZY ITEM-SETS AND FUZZY ASSOCIATION RULES

Let $I = \{I_1, I_2, ..., \mathrm{I_m}\}$ be a set of fuzzy items. Let D, the task-relevant fuzzy data, be a set of database transactions where each transaction T is a set of fuzzy items such that $T \subseteq I$. Let A be a set of fuzzy items. A transaction T is said to contain A if and only if $A \subseteq T$. A fuzzy association rule is an implication of the form $A \Rightarrow B$, where $A \subset I, B \subset I$ and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with fuzzy support s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability $P(A \cup B)$. The rule $A \Rightarrow B$ has fuzzy confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability,

$$P(B/A). \quad \textit{fuzzysupport} \quad (A \Rightarrow B) = P(A \cup B) \; \textit{fuzzy confidence} \; (A \Rightarrow B) = P(B/A) \tag{1}$$

A set of items is referred to as an item-set. An item-set that contains k items is called a k-item-set. The occurrence frequency of an item-set is the number of transactions that contain the item-set. The fuzzy item-set support defined in Equation (1) is called relative support, whereas the occurrence frequency is called the absolute support. If the relative support of an item-set I, satisfies a pre-specified minimum support threshold, then I is a frequent fuzzy item-set. From (1) we have:

$$\begin{aligned} &\textit{fuzzy confidence} \; (A \Rightarrow B) = P(B/A) \\ &= \frac{\sup port(A \cup B)}{\sup port(A)} \\ &= \frac{\sup port\_count(A \cup B)}{\sup port\_count(A)} \end{aligned} \tag{2}$$

Equation (2) shows that the fuzzy confidence of rule $A \Rightarrow B$ can be easily derived from the support counts of A and $A \cup B$. Once the support counts of A, B and $A \cup B$ are found, it is straightforward to derive the corresponding association rules $A \Rightarrow B$ and $B \Rightarrow A$, and check whether they are strong. Thus the problem of mining fuzzy association rules can be reduced to that of mining frequent fuzzy item-sets. A fuzzy item-set X is closed in a fuzzy data set S if there exists no proper super-item set Y such that Y has the same support count as X in S. An item-set X is a closed frequent fuzzy item-set in set S if X is both closed and frequent in S.

FUZZY ASSOCIATION RULES

The fuzzy item-sets which frequently occur together in large databases are found using fuzzy association rules (Fu et al., 1998). All the methods used for mining fuzzy association rules are based upon a support-confidence framework where fuzzy support and fuzzy confidence are used to identify the fuzzy association rules. Let $F = \{f_1, f_2, ..., f_m\}$ be a set of fuzzy items, $T = \{t_1, t_2, ..., t_n\}$ be a set of fuzzy records, and each fuzzy record t_i is represented as a vector with m values, $(f_1(t_i), f_2(t_i), ..., f_m(t_i))$, where $f_j(t_i)$ is the degree that f_j appears in record t_i, $f_j(t_i) \in [0,1]$. Then a fuzzy association rule is defined as an implication form such as $F_X \Rightarrow F_Y$, where $F_X \subset F, F_Y \subset F$ are two fuzzy item-sets.

The fuzzy association rule $F_X \Rightarrow F_Y$ holds in T with the fuzzy support ($\textit{fsupp}(\{F_X, F_Y\})$) and the fuzzy confidence ($\textit{fconf}$ $(F_X \Rightarrow F_Y)$). The fuzzy support and fuzzy confidence are given as follows:

$$\textit{fsupp}(\{F_X, F_Y\}) = \frac{\sum_{i=1}^{n} \min(f_j(t_i) / f_j \in \{F_X, F_Y\})}{n} \tag{3}$$

$$\textit{fconf}\,(F_X \Rightarrow F_Y) = \frac{\textit{fsupp}\,(\{F_X, F_Y\})}{\textit{fsupp}\,(\{F_X\})} \tag{4}$$

If the $\textit{fsupp}(\{F_X, F_Y\})$ is greater than or equal to a predefined threshold, minimal fuzzy support (s_f), and the $\textit{fconf}$ $(F_X \Rightarrow F_Y)$ is also greater than or equal to a predefined threshold, minimum fuzzy confidence (c_f),

then $F_X \Rightarrow F_Y$ is considered as an interesting fuzzy association rule, and it means that the presence of the fuzzy item-set F_X in a record can imply the presence of the fuzzy item sets F_Y in the same record. If a practical situation is considered where a fuzzy item-set almost occurs in all fuzzy records, then according to the aforementioned framework, many fuzzy association rules can be identified. Interestingly, the presence of this fuzzy item-set does not necessarily imply the presence of other fuzzy item-sets which are also included in these fuzzy association rules. Hence there is an urgent need for analysing the relationships between fuzzy item-sets. Fuzzy correlation analysis is used to determine the linear relationship between any two fuzzy item-sets.

FROM FUZZY ASSOCIATION ANALYSIS TO FUZZY CORRELATION ANALYSIS

As it is seen, the fuzzy support and fuzzy confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, a fuzzy correlation measure Bustince and Burillo (1995), Hong and Hwang (1995) can be used to augment the fuzzy support-confidence framework for fuzzy association rules. This leads to the fuzzy correlation rules of the form $A \Rightarrow B$. Hence the correlation between the fuzzy item-sets A and B becomes necessary. There are many different fuzzy correlation measures from which to choose (Yu, 1993; Kao and Liu, 2002).

In this paper, the method proposed C. Kao and S.T. Liu is followed due to the advantages it has, compared to the earlier proposed fuzzy correlation methods (Chiang and. Lin, 1999). This method works well especially in an environment of fuzzy numbers.

When all the observations are fuzzy, equation for correlation of fuzzy numbers is given by

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \sum_{i=1}^{n}\frac{X_i}{n}\right)\left(Y_i - \sum_{i=1}^{n}\frac{Y_i}{n}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \sum_{i=1}^{n}\frac{X_i}{n}\right)^2 \sum_{i=1}^{n}\left(Y_i - \sum_{i=1}^{n}\frac{Y_i}{n}\right)^2}} \quad (5)$$

Since it is difficult to derive membership function for $\rho_{X,Y}$ for (5) directly, we rely on Zadeh's extension principle [15] (Zadeh, 1978) which says

$$\mu_{\rho_{X,Y}}(\rho) = \sup_{X,Y} \min\left\{\mu_{X_i}(x_i), \mu_{Y_i}(y_i) \forall i / \rho = \rho_{X,Y}\right\}$$

Equation (5) is used to calculate the correlation coefficient between the fuzzy numbers and the value computed from (5) lies between the interval [-1,1]. According to the aforementioned proposed method we can obtain the strength and type of the linear relationship between two fuzzy item-sets. Hence the fuzzy correlation analysis is of great use in mining only the interesting fuzzy correlation rules.

MINING FUZZY ASSOCIATION AND FUZZY CORRELATION RULES

Mining fuzzy association rules is better done by finding frequent fuzzy item-sets using candidate generation method (Han and Kamber, 2001). Apriori is a seminal algorithm proposed for mining frequent fuzzy item-sets. The algorithm uses prior knowledge of frequent fuzzy item-set properties. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1) –item sets. First, the set of 1-itemsets is found by scanning the fuzzy database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by L1. Next L1 is used to find L2, the set of frequent fuzzy 2-itemset, which is used to find L3, and so on, until no more frequent fuzzy k-item-sets can be found. The finding of each Lk requires one full scan of the database. The algorithm consists of two steps, namely (i) the join step and (ii) the prune step, for candidate generation.

CASE STUDY

A higher secondary school in Trichy keeps the record of the revision exams conducted for the class of 12 students. The school has 650 students in the higher secondary levels, both boys and girls. Repeated exams are conducted prior to the final public exams. The management is interested in performing the market basket analysis with regards to the subjects the students like the most and choose the same in the exams, so that special concentration can be given to those subjects and thereby improving the results of school. A sample of 15 repeated revision exams of the full portions in the subject of Mathematics is taken for the study. The subject has 10 different chapters, connected to each other in one or many means. Association and correlation rules are applied for these different chapters to discover interesting relationships. Table 1 is the fuzzy scores of the overall performance for each individual chapter. In the table f1, f2... f10 are the 10 chapters in Mathematics and t1, t2... t15 are the transactions,(revision exams).

Table 2 represents the fuzzy support of the 10 chapters of the subject Mathematics and Figure 1 gives a clear picture of the same.

Here in the study the fuzzy support is fixed to 0.20; the fuzzy confidence is fixed to 0.85; the minimal fuzzy correlation coefficient is fixed as r_f = 0.20; the level of

Table 1. Fuzzy scores for the performance.

F / T	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
t1	0.1	0.2	0.1	0.7	0.6	0.1	0.7	0.7	0.5	0.8
t2	0.1	0.3	0.2	0.8	0.3	0.7	0.9	0.8	0.2	0.7
t3	0.2	0.8	0.7	0.9	0.6	0.1	0.2	0.3	0.8	0.5
t4	0.3	0.2	0.4	0.8	0.8	0.3	0.8	0.9	0.1	0.7
t5	0.4	0.3	0.5	0.7	0.9	0.7	0.9	0.8	0.3	0.8
t6	0.2	0.9	0.3	0.6	0.7	0.2	0.1	0.1	0.9	0.3
t7	0.9	0.1	0.4	0.8	0.2	0.3	0.9	0.9	0.4	0.8
t8	0.3	0.4	0.1	0.7	0.8	0.2	0.1	0.2	0.8	0.4
t9	0.8	0.9	0.2	0.8	0.9	0.2	0.7	0.4	0.1	0.6
t10	0.2	0.9	0.1	0.7	0.1	0.9	0.7	0.9	0.4	0.5
t11	0.1	0.4	0.2	0.9	0.2	0.1	0.9	0.8	0.9	0.8
t12	0.3	0.1	0.3	0.8	0.7	0.2	0.1	0.2	0.8	0.2
t13	0.1	0.9	0.2	0.8	0.4	0.9	0.3	0.1	0.2	0.5
t14	0.1	0.2	0.1	0.9	0.2	0.7	0.9	0.8	0.1	0.7
t15	0.4	0.1	0.2	0.7	0.1	0.3	0.8	0.9	0.7	0.6

Table 2. The fuzzy support of the fuzzy items.

F	Fsupport
f1	0.30
f2	0.45
f3	0.27
f4	0.78
f5	0.50
f6	0.40
f7	0.60
f8	0.59
f9	0.48
f10	0.59

significance for t-distribution is set to 0.01, and the distribution value is 2.65. The calculation of r value in table is done using the fuzzy correlation approach given by Liu and Kao (2002). The calculation of t value in table is done using Arnold (1990) test statistic given by

$$t = \frac{r_{A,B} - r_f}{\sqrt{\frac{1 - r_{A,B}}{n-2}}}$$

C2 consists of 45 fuzzy 2-itemsets, but only those which satisfy the minimum threshold level as earlier discussed is given Table 3. Figure 2 shows the comparison of the items of C2.

From all the fuzzy 2-itemset combinations an element whose Fsupport is greater than or equal to 0.20 and t-value greater then equal to 2.65 is considered as an element of L2. L2={ ({f7,f8},{f7,f10},{f8,f10})}. Now C3 is generated by joining L2 with L2. Table 4 gives a clear picture of the members of C3.

In Table 4 all elements of C3 satisfy the minimum threshold levels. Hence all the elements of C3 are elements of L3. Thus L3 = C3. Now the next C4 cannot be generated by joining L3 with L3, and here the mining procedure stops. By this process some 12 candidate fuzzy correlation rules can be generated out of which only 7 rules are interesting because some do not satisfy the minimum threshold level of confidence 0.85.

$$\{f7\} \rightarrow \{f8\}$$
$$\{f8\} \rightarrow \{f7\}$$
$$\{f7\} \rightarrow \{f10\}$$
$$\{f10\} \rightarrow \{f7\}$$
$$\{f8, f10\} \rightarrow \{f7\}$$
$$\{f7, f10\} \rightarrow \{f8\}$$
$$\{f7, f8\} \rightarrow \{f10\}$$

CONCLUSION

In this paper the method for mining fuzzy correlation rules is based on the fuzzy measures for correlation coefficient of fuzzy numbers. The method of correlation analysis proposed in the fuzzy correlation techniques gives some interesting observations of some 7 fuzzy correlation rules. Here in our study f7, f8 and f10 are the chapters namely Integral-calculus, differential equations and probability distributions. Through this study it is seen that the

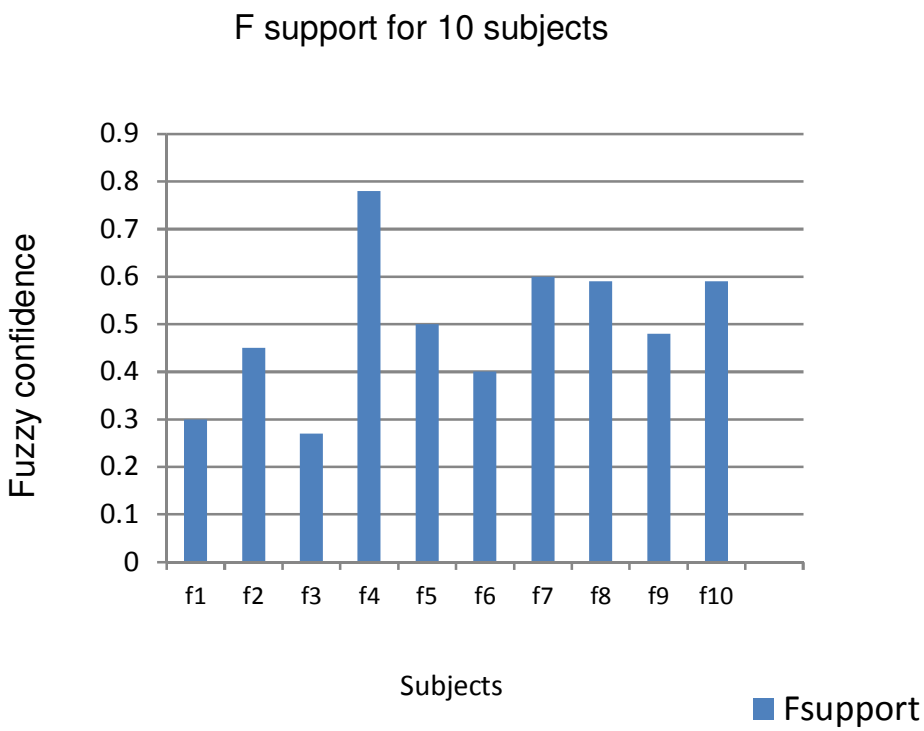

Figure 1. Fuzzy support for the 10 chapters.

Table 3. Candidate Generation of 2-itemsets.

C2	Fsupport	r	t
{f7}{f8}	0.54	0.9156	6.2662
{f7}{f10}	0.51	0.8863	5.7001
{f8}{f10}	0.50	0.7708	3.2303

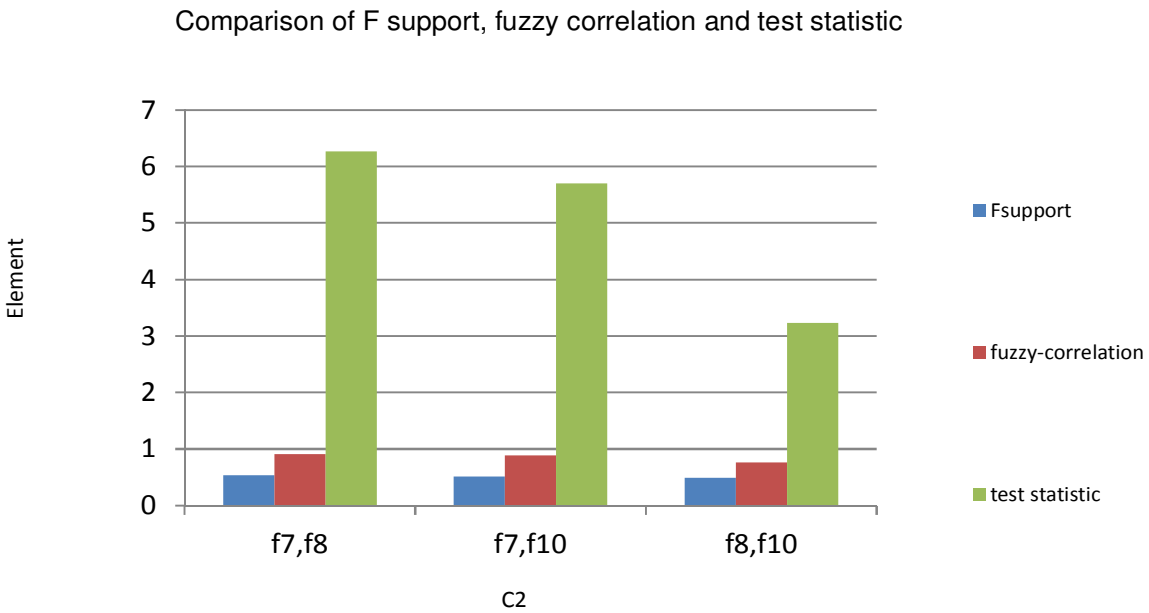

Figure 2. Comparision of F support, r and t of C2.

Table 4. Candidate Generation of 3-itemsets.

C3	Fsupport	r	t
{f7},{f8,f10}	0.48	0.9337	9.2522
{f8},{f7,f10}	0.48	0.8743	5.0076
{f10}{f7,f8}	0.48	0.8523	4.4955

students often prefer questions from these chapters in their examinations.

REFERENCES

Agrawal R, Imielinski T, Swami A (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington D.C., May, pp. 207-216.

Agrawal R, Strikant R (1994). Fast algorithms formining association rules, Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, Sept., pp. 487-499.

Arnold SF (1990). Mathematical Statistics, Prentice-Hall, New Jersey.

Au W-H, Chan KCC (1998). An effective algorithm for discovering fuzzy rules in relational databases. Proceedings of the IEEE World Congress on Computational Intelligence, pp. 1314–1319.

Bustince H, Burillo P (1995). Correlation of interval-valued intuitionistic fuzzy sets. Fuzzy sets Syst., 74: 237-244.

Chan KCC, Wong AKC (1997). Mining Fuzzy Association Rules. Proceedings of the Sixth International Conference on Information and Knowledge Management, Las Vegas, Nevada, United States, Nov., pp. 209-215.

Chiang DA, Lin NP (1999). Correlation of Fuzzy Sets. Fuzzy Sets Syst., 102: 221-226.

Dubois D, Hullermeier E, Prade H (2003). A Note on Quality Measures for Fuzzy Association Rules. Proceedings of the 10thInternational Fuzzy Systems Association World Congress (IFSA-03), Lecture Notes in Artificial Intelligence 2715, Springer-Verlag, pp. 346-353.

Dunham MH (2003). Data mining, Introductory and Advanced Topics. Pearson Education Inc.

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

Han J, Kamber M (2001). Data mining: Concepts and Techniques. Academic Press.

Hong DH, Hwang SY (1995). Correlation of intuitionistic fuzzy sets in probability spaces. Fuzzy Sets Syst., 75: 77-81.

Kao C, Liu ST (2002). Fuzzy measures for correlation coefficient of fuzzy numbers. Fuzzy Sets Syst., 128: 267-275.

Yu C (1993). Correlation of fuzzy numbers. Fuzzy Sets Syst., 55: 303-307.

Zadeh LA (1978). Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst., 1: 3-28.