2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

CPDA Based Fuzzy Association Rules for Learning Achievement Mining

Jr-Shian Chen[1+], Hung-Lieh Chou[2,3], Ching-Hsue Cheng[2], Jen-Ya Wang[1]

[1]Department of Computer Science and Information Management, HUNGKUANG University

[2] Department of Information Management, National Yunlin University of Science and Technology

[3] Department of Computer Center, St. Joseph's Hospita

Abstract. This paper proposes a fusion model to reinforce fuzzy association rules, which contains two main procedures: (1) employing the cumulative probability distribution approach (CPDA) to partition the universe of discourse and build membership functions; and (2) using the AprioriTid mining algorithm to extract fuzzy association rules. The proposed model is more objective and reasonable in determining the universe of discourse and membership functions with other fuzzy association rules.

Keywords: cumulative probability distribution approach (CPDA), fuzzy association rule, AprioriTid

1. Introduction

Previous studies [4, 5] often partitioned the length of interval in equal-length and ignored the distribution characteristics of datasets. Therefore, this paper focuses on improving the persuasiveness in determining the universe of discourse and membership functions of fuzzy association rules. In empirical case study, we use an exemplary dataset to be our simulation data, which contains the learning achievement data of 10 students.

The remaining content of this paper is organized as follows. Related work is described in section 2. Section 3 presents our proposed approach. Finally, conclusions of this paper are stated in the Section 4.

2. Related Work

In this section, we will briefly discuss the following research: fuzzy numbers, fuzzy association rules, and our prior research cumulative probability distribution approach (CPDA) [1].

2.1. Cumulative probability distribution approach

Our prior work CPDA [1] partitions the universe of discourse and builds membership functions, which is based on the inverse of the normal cumulative distribution function. In probability theory, the inverse of the normal cumulative distribution function (CDF) is modeled by two parameters μ and σ for a given probability p. The CDF is defined as follows:

$$x = F^{-1}(p \mid \mu, \sigma) = \{x : F(x \mid \mu, \sigma) = p\}, \tag{1}$$

where

$$p = F(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty^{e}}^{x} \frac{-(t-\mu)^2}{2\sigma^2} dt, \tag{2}$$

and μ, σ denote the mean and the standard deviation, respectively.

[+] Corresponding author. Tel.: + 886-4-26318652x5420; fax: +886-4-26521921.
E-mail address: jschen@sunrise.hk.edu.tw.

2.2. Fuzzy numbers

Zadeh (1965) first introduced the concept of fuzzy set for modeling the vagueness type of uncertainty [2]. A fuzzy set $\tilde{A}$ defined on the universe X is characterized by a membership function $\mu_{\tilde{A}} : x \rightarrow [0,1]$.

When describing imprecise numerical quantities, one should capture intuitive concepts of approximate numbers or intervals such as "approximately m." A fuzzy number must have a unique modal value "m", convex and piecewise continuous. A common approach is to limit the shape of membership functions defined by LR-type fuzzy numbers. A special case of LR-type fuzzy numbers TFN is defined by a triplet, denoted by $A \rightarrow (a,b,c)$. The graph of a typical TFN is shown in Fig. 1.

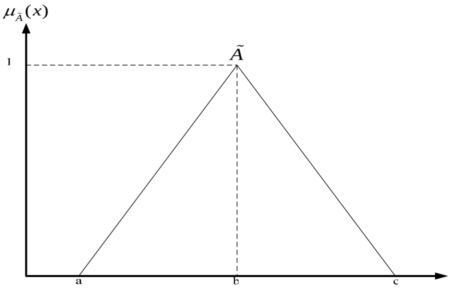

Fig. 1: Triangular fuzzy numbers.

2.3. Fuzzy association rules

Agrawal et al. [3] introduced Apriori approach to explore behaviors in market basket data, those transactions are all binary values. However, any real-world database may also contain quantitative attributes. To overcome there shortcomings, fuzzy association rule mining approaches were proposed to handle both the quantitative and categorical attributes. Fuzzy association rules-based methods transform quantitative values into a fuzzy set for each attribute; and use the membership degree operations to find rules. These rules can be expressed like the one: If age is young, then salary is low. Detailed overviews of fuzzy association rules mining methods can be found in [4, 5].

3. The Proposed Approach

In this section, the fusion model is proposed to improve the persuasiveness in determining the universe of discourse and membership functions of fuzzy association rules. Firstly, we build the membership function based on CPDA for each conditional attribute. Secondly, we use the AprioriTid mining algorithm to extract fuzzy association rules. An exemplary dataset [4] used throughout this paper is shown in Table 1. The dataset contains 10 instances, which is characterized by the following attributes: (I) statistics (denoted ST), (II) database (denoted DB), (III) object-oriented programming (denoted OOP), (IV) data structure (denoted DS), and (V) management information system (denoted MIS). All these attributes are numerical values. The proposed model is introduced in detail as follows:

Table 1. The exemplary data set.

Student No.	ST	DB	OOP	DS	MIS
1	86	77	86	71	68
2	61	79	89	77	80
3	84	89	86	79	89
4	73	86	79	84	62
5	70	89	87	72	79
6	65	77	86	61	87
7	67	87	75	71	80
8	86	63	64	84	86
9	75	65	79	87	88
10	79	63	63	85	89

Step 1: Partition continuous attributes by CPDA.

The CPDA is used to discrete the dataset and define the fuzzy numbers. This model partitions the universe of discourse based on cumulative probability distribution. The results are shown in Table 2.

Table 2. The fuzzy number for each attribute.

Attribute	Low	Middle	High
ST	(48.22,58.65,69.09)	(62.23,71.17,80.11)	(74.60,86.69,98.78)
DB	(47.95,59.48,71.02)	(62.94,73.46,83.98)	(77.50,88.75,100.00)
OOP	(49.53,61.56,73.60)	(66.37,75.79,85.20)	(79.40,89.70,100.00)
DS	(49.20,60.61,72.02)	(65.68,73.93,82.18)	(77.10,87.95,98.80)
MIS	(48.81,61.96,75.12)	(68.04,77.26,86.48)	(80.80,90.40,100.00)

Step 2: Build membership function.

The membership function is built by employing triangular fuzzy number and the triangular fuzzy number model of Table 2 is shown in Fig. 2. The membership function as shown in Fig. 2 can be established based on CPDA.

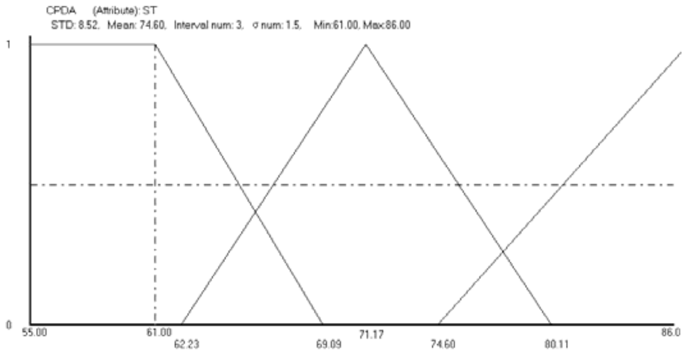

Fig. 2: Membership functions of ST attribute.

Step 3: Fuzzify the continuous data.

According to the membership function in step 2, the degree of membership for each datum is calculated. Table 3 shows the result of the students' achievement dataset.

Table 3. The degree of membership for each datum.

No.	ST			DB			OOP			DS			MIS		
	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
1	0.00	0.00	0.94	0.00	0.66	0.00	0.00	0.00	0.64	0.09	0.64	0.00	0.54	0.00	0.00
2	0.78	0.00	0.00	0.00	0.47	0.13	0.00	0.00	0.93	0.00	0.63	0.00	0.00	0.70	0.00
3	0.00	0.00	0.78	0.00	0.00	1.00	0.00	0.00	0.64	0.00	0.39	0.18	0.00	0.00	0.85
4	0.00	0.80	0.00	0.00	0.00	0.76	0.00	0.66	0.00	0.00	0.00	0.64	1.00	0.00	0.00
5	0.00	0.87	0.00	0.00	0.00	1.00	0.00	0.00	0.74	0.00	0.77	0.00	0.00	0.81	0.00
6	0.39	0.31	0.00	0.00	0.66	0.00	0.00	0.00	0.64	0.97	0.00	0.00	0.00	0.00	0.65
7	0.20	0.53	0.00	0.00	0.00	0.84	0.00	0.92	0.00	0.09	0.64	0.00	0.00	0.70	0.00
8	0.00	0.00	0.94	0.70	0.01	0.00	0.80	0.00	0.00	0.00	0.00	0.64	0.00	0.05	0.54
9	0.00	0.57	0.03	0.52	0.20	0.00	0.00	0.66	0.00	0.00	0.00	0.91	0.00	0.00	0.75
10	0.00	0.12	0.36	0.70	0.01	0.00	0.88	0.00	0.00	0.00	0.00	0.73	0.00	0.00	0.85

Step 4: Predefine the minimum support value and confidence threshold.

Users need to predefine the minimum support value and the confidence threshold. The minimum support value is set to 2.0 and the confidence threshold is set at 0.7 in this case.

Step 5: Generate the candidate itemset C_1.

Taking the linguistic value ST.Low as an example, the scalar cardinality will be (0.78 + 0.39 + 0.20) = 1.37. The C_1 candidate itemset for this example is shown as follows:

{(ST.low,1.37), (ST.middle,3.20), (ST.high,3.05), (DB.low,1.92), (DB.middle,2.01), (DB.high,3.73), (OOP.low,1.68), (OOP.middle,2.24), (OOP.high,3.59), (DS.low,1.15), (DS.middle,3.07), (DS.high,3.10), (MIS.low,1.54), (MIS.middle,2.26), (MIS.high,3.64)}

Step 6: Generate the L_1 large itemset.

The L_1 large itemset rides on the largest count value for each attribute and is equal to or greater than the minimum support value. In this exemplary dataset, L_1 can be denoting as follows:

{(ST.middle,3.20), (DB.high, 3.73), (OOP.high,3.59), (DS.high, 3.10), (MIS.high, 3.64)}.

Step 7: Generate the candidate itemset C_{1+i} from L_i.

The candidate itemset is generated from L_i. Here, comparison operation is used to return the lesser membership degree of a newly formed candidate itemset. In the example below, we compare the linguistic values ST.middle with DB.high that is transformed into ST.middle,DB.high. The result is shown in Table 4.

Table 4. The linguistic value *(ST.middle,DB.high).*

No.	*ST.middle*	*DB.high*	*(ST.middle,DB.high)*
1	0.00	0.00	0.00
2	0.00	0.13	0.00
3	0.00	1.00	0.00
4	0.80	0.76	0.76
5	0.87	1.00	0.87
6	0.31	0.00	0.00
7	0.53	0.84	0.53
8	0.00	0.00	0.00
9	0.57	0.00	0.00
10	0.12	0.00	0.00

Taking the linguistic value (ST.middle,DB.high) as an example, the scalar cardinality will be (0.76 + 0.87 + 0.53) = 2.16. The C_2 candidate itemset for this exemplary dataset is shown in Table 5.

Table 5. The C_2 candidate itemset.

Itemset	Count
ST.middle,DB.high	2.16
ST.middle,OOP.high	1.05
ST.middle,DS.high	1.33
ST.middle,MIS.high	1.00
DB.high,OOP.high	1.51
DB.high,DS.high	0.82
DB.high,MIS.high	0.85
OOP.high,DS.high	0.18
OOP.high,MIS.high	1.28
DS.high,MIS.high	2.20

Step 8: Generate the L_{1+i} large itemset from candidate itemset C_{1+i}.

The L_2 large itemset rides on the count value, which is equal to or greater than the minimum support value. The L_2 large itemset for this exemplary dataset is shown as follows:

{((ST.middle,DB.high),2.16),(((DS.high,MIS.high)),2.20)}

Table 6. The linguistic value *(ST.middle,DB.high,DS.high).*

No.	*ST.middle*	*DB.high*	*DS.high*	*(ST.middle,DB.high,DS.high)*
1	0.00	0.00	0.00	0.00
2	0.00	0.13	0.00	0.00
3	0.00	1.00	0.18	0.00
4	0.80	0.76	0.64	0.64
5	0.87	1.00	0.00	0.00
6	0.31	0.00	0.00	0.00
7	0.53	0.84	0.00	0.00
8	0.00	0.00	0.64	0.00
9	0.57	0.00	0.91	0.00
10	0.12	0.00	0.73	0.00

Step 9: Repeat the steps 7 and 8 until the L_{1+i} large itemset is null.

In the example below, we generate the candidate itemsets C_3. The count value of linguistic value (ST.middle,DB.high,DS.high) is described in Table 6. The C_3 candidate itemset is listed in Table 7. All of the count values in C_3 are smaller than the minimum support value. Therefore, the L_3 large itemset is null.

Table 7. The C_3 candidate itemset..

Itemset	Count
ST.middle,DB.high,DS.high	0.64
ST.middle,DB.hig,MIS.high,	0.00
ST.middle DS.high,MIS.high	0.69
DB.high,DS.high,MIS.high	0.18

Step 10: Construct the fuzzy association rules.

Construct the fuzzy association rules from L_2. In the following example, $ST.middle \Rightarrow DB.high$, its confidence value is calculated as follows: $\frac{(ST.middle, DB.high)}{ST.middle} = \frac{2.16}{3.20} = 0.68$. The fuzzy association rules are listed in Table 8.

Table 8. The fuzzy association rules.

fuzzy association rules	confidence value
$ST.middle \Rightarrow DB.high$	0.68
$DB.high \Rightarrow ST.middle$	0.58
$DS.high \Rightarrow MIS.high$	0.71
$MIS.high \Rightarrow DS.high$	0.60

Now we can check whether the confidence values of the above association rules (listed in Table 8) are larger than or equal to the predefined confidence threshold. The final resulting rules are thus obtained: If the score of data structure is high, then the score of management information system is high.

4. Conclusion

This paper proposes a fusion model to improve the persuasiveness in determining the universe of discourse and membership functions, which combines AprioriTid and our prior work CPDA. The proposed model is more objective and reasonable in determining the universe of discourse and membership functions with other fuzzy association rules.

5. Acknowledgements

This work was funded by the National Science Council of the Republic of China under contract NSC 97-2511-S-241-007-MY3.

6. References (This is “Header 1” style)

[1] Teoh, H.J., Cheng, C.H., Chu, H.H. and Chen, J.S., Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets, *Data & Knowledge Engineering*, 2008, 67(1): 103-117

[2] Ross, T.J., *Fuzzy logic with engineering applications*, International edition, McGraw-Hill, USA. (2000)

[3] Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases. *In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD-93)*, Washington, DC, United States, 1993, 207–216

[4] Hong, T.P., Kuo, C.S. and Wang, S.L., A fuzzy AprioriTid mining algorithm with reduced computational time. *Applied Soft Computing*, 2004, 5(1): 1-10.

[5] Khan, M. S., Muyeba, M. K., and Coenen, F.: Weighted Association Rule Mining from Binary and Fuzzy Data, *presented at Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, 8th Industrial Conference, ICDM 2008*, Leipzig, Germany, 2008, 200-212.