A Study of Fuzzy Quantitative Items Based on Weighted Association Rules Mining

Tianqi Yang[a], Chengjun Li[b]
Department of Computer Science, Jinan University, Guangzhou, China
[a]e-mail: tytq@jnu.edu.cn, [b]e-mail: 340345085@qq.com

***Abstract*—The weighted association rules mining is more significative than traditional association rules mining in practice. Allowing for the impact of the number and weight of property on association rules, this paper presents a new method of mining weighted association rules, which can hold the " downward closed property " by using an improved model of weighted support measurements in the weighted setting. Compared to some generalized weighted association rules mining, it proves that the method can quickly and efficiently mine important association rules.**

Keywords-weighted association rules; weighted support; quantitative itemsets

I. INTRODUCTION

Data mining and knowledge discovery in databases is an interesting research area only developed in the last fifteen years. Association rule mining aims to explore large transaction databases for association rules, which may reveal the implicit relationships among the data attributes. It can be divided into boolean association rules and quantitative attribute association rules. Agrawal first proposed boolean association rules in 1993, and then proposed the classic Apriori and Apriori TID algorithm[1], In general the actual database is boolean and quantitative attribute mixed.Association rule mining[2] is a popular data mining technique because of its wide application in marketing and retail communities as well as other more diverse fields [3][15].

Association Rule Ming（ARM）is an important and well established data mining topic.The problem of association rules (AR) can be expressed as $X \Rightarrow Y$, which means that a data record that contains the set of items X is likely to contain items Y as well. If X or Y in AR is a set of fuzzy sets, then we call this kind of association rules fuzzy ARs.Recently, various efforts have been made to develop and improve theory or applications of fuzzy ARs.For example,[3] introduces two kind of fuzzy extensions to classical ARs, One extension is that, in generalized association rules, the taxonomies concerned is fuzzy. The other extension is that managers are likely to use fuzzy linguistic expressions when referring to decision rules. [4] focus on positive and negative fuzzy association rules mining. Fuzzy rules also can be obtained in quantitative rules by replacing intervals by fuzzy sets.

Fuzzy association rule mining is proposed to discover other knowledge from a large database. It is especially useful to consider the quantities involved in the user's request. If generic association rules mining can discover R1: Beer->Potato, a fuzzy association rule can discover R2 :(Beer, 1 bottle)->(Potato,150 gram),Such merit presents more detail and a more reliable exploration of the association rules than a generic association rule. And [4] focus on positive and negative fuzzy association rules mining. Fuzzy rules also can be obtained in quantitative rules by replacing intervals by fuzzy sets. [7] Introduces the problem of mining weighted quantitative association rules based on fuzzy approach. Using the fuzzy set concept, the discovered rules are more understandable to a human.[8] assumes an understanding of the semantic meaning of a fuzzy rule and develops a systematic approach to the assessment of fuzzy association rules.[9] proposed a novel algorithm to avoid the loss of semantic information due to the partition of quantitative values.[10] defines a fuzzy data cube, which facilitates for handling quantitative values of dimensional attributes, and hence allows for mining fuzzy association rules at different levels.[11] addresses the integration of fuzziness with On-Line Analytical Processing(OLAP) based association rules mining. It contributes to the ongoing research on multidimensional online association rules mining by proposing a general architecture that utilizes a fuzzy data cube for knowledge discovery. [12] Proposes the problem of mining weighted generalized fuzzy association rules with fuzzy taxonomies. It is an extension of the generalized fuzzy association rules with fuzzy taxonomies problem. [13] addresses the issue of invalidation of downward closure property (DCP) in weighted association rule mining where each item is assigned a weight according to its significance w.r.t some user defined criteria and generalizes the weighted association rule mining problem for databases with binary and quantitative attributes with weighted settings.[14] focuses on the notion of fuzzy association rules form a collection of fuzzy sets and deals with the fuzziness based upon fuzzy taxonomies that reflect partial belongings among item sets, as well as upon the extended settings for the degree of support and the degree of confidence.[15] proposes and implements an intelligent information system that is based on both a wired and a wireless telephone network-based speech Web using fuzzy association rule mining.[16] introduces a new measure w-support, which does not require preassigned weights. It takes the quality of transactions into consideration using link-based models.

II. PROBLEM DEFINITION

In this section formal definitions are presented to define quantitative attributes, The FARM concept and the normalization process for Fuzzy Transactions (*FT*).

A. Terms and definitions

Fuzzy Association Rules A Fuzzy AR [18] is an implication of the form: if <*X*, *A*> then <*Y*, *B*>, where X and Y are disjoint item sets and A and B contain the fuzzy sets associated with the corresponding attributes in X and Y. As in the binary association rule, "X is A" is called the antecedent of the rule while "Y is B" is called the consequent of the rule.

Fuzzy Frequent Item sets An itemset<*X*, *A*> is called a frequent itemset if its fuzzy support value is greater than or equal to a minimum support threshold.

We use the discovered frequent itemsets to generate all possible rules. If the union of antecedent <*X*, *A*> and consequent <*Y*, *B*> has enough support and the rule has high condence, this rule will be considered as interesting. When we obtain a frequent itemset <*Z*, *C*>, we want to generate fuzzy association rules of the form, "If X is A then Y is B", where $X \subset Z$, $Y = Z - X$, $A \subset C$ and $B = C - A$. Having the frequent itemset, we know its support as well as the fact that all of its subsets will be also frequent.

Fuzzy Support and Confidence if X is a set of attribute-fuzzy set label pairs. A record t_i satisfies Ex. $X \subseteq t_i$ .The individual vote per record is found by multiplying the membership degree with an attribute-fuzzy set pair $[i[l]] \in X$:

vote for t_i satisfying $X = \prod_{(\forall [i[l]] \in X)} t_i[i[l]]$ .so we have

$$FS(X) = \frac{\sum_{i=1}^{n} \prod_{(\forall [i[l]] \in X)} t_i[i[l]]}{n} \qquad (1)$$

Frequent attribute sets with fuzzy support above the specified threshold are used to generate all possible rules. A fuzzy AR derived from a fuzzy frequent attribute set C is of the form:

$A \rightarrow B$, where A and B are disjoint subsets such that $C = A \cup B$. Fuzzy Confidence (*FC*) is calculated in the same manner that confidence is calculated in classical ARM:

$$FC(A \rightarrow B) = \frac{FS(A \cup B)}{FS(A)}. \qquad (2)$$

Downward Closure Property (DCP) In classical ARM algorithm, it is assumed that if the itemset is large, then all its subsets should also be large, a principle called downward closure property (DCP) or anti-monotonic property of itemsets. For example, in standard ARM using DCP, it states that if {AB} and {BC} are not frequent, then {ABC} and {BCD} can not be frequent, consequently their supersets are of no value as they will contain non-frequent itemsets. This helps the algorithm to generate large itemsets of increasing size by adding items to itemsets that are already large.

B. Improved Fuzzy Weighted Association Rule Mining

A fuzzy dataset *D* consists of fuzzy transactions $T=\{t_1, t_2, \ldots\ldots, t_n\}$ with fuzzy sets associated with each item in $I=\{i_1, i_2, \ldots, i_{|I|}\}$,which is identified by a set of linguistic labels $L=\{l_1, l_2, \ldots, l_{|L|}\}$,for example *L*={Small, Medium, Large},We assign a weight *w* to each *l* in *L*.

Associated with *I*. Each attribute $t_i[i_j]$ is associated with several fuzzy sets The degree of association is given by a membership degree in the range [0, 1], which indicates the correspondence between the value of a given $t_i[i_j]$ and the set of fuzzy linguistic labels. The "k^{th}" weighted fuzzy set for the "j^{th}" item in the "i^{th}" fuzzy transaction is given by $t_i[i_j[l_k[w]]]$ .Thus each label l_k for item i_j would have associated with it a weight, i.e. a pair $([i[l]], w)$ is called a weighted item where $[i[l]] \in L$ is a label associated with I and $w \in W$ is weight associated with label *l*.

Fuzzy Item Weight FIW is a value attached with each fuzzy set. It is a non-negative real number value range [0, 1] to list some degree of importance eight of a fuzzy set for an item i_j is denoted as $i_j[l_k[w]]$.

Fuzzy Itemset Transaction Weight FITW is the aggregated weights of all the fuzzy sets associated to items in the itemset present in a single transaction. Fuzzy Itemset transaction weight for an itemset (*X*, *A*) is calculated as:

Vote for t_i satisfying

$$X = \frac{\min(X)}{\max(X)} \prod_{k=1}^{|L|} {}_{(\forall [i[l[w]]] \in X)} t_i[i_j[l_k[w]]] \qquad (3)$$

TABLE I. FUZZY TRANSACTIONAL DATABASE

TID	X		Y	
	Small	Large	Small	Large
T1	0.9	0.1	0.4	0.6
T2	0.2	0.8	0.5	0.5
T3	1.0	0.0	0.3	0.7
T4	0.6	0.4	0.2	0.8

TABLE II. FUZZY ITEMS WITH WEIGHTS

Fuzzy Items i[l]	Weights(IW)
(X,Small)	0.8
(X,Large)	0.5
(Y,Small)	0.6
(Y,Large)	0.3

C. Weighted Downward Closure Property

In a classical Apriori algorithm, DCP helps algorithm to generate large itemsets of increasing size by adding items to itemsets that are already large. But in the weighted ARM case where each item is assigned a weight, he DCP does not hold. Because of the weighted support, n item set may be

large even though some of its subsets are not large. Due to induct the weights, one weighted ARM algorithm [5 7] can not content the DCP .Now we argue that the DCP with quantitative data can be validated using the proposed approach. We also briefly prove that the DCP is always valid in the proposed method. He following lemma applies to both Boolean and quantities data and is stated as:

Lemma

If an itemset is not frequent then its superset can not be frequent and *FS* (subset) $\geq$ *FS*(superset) is true.

III. IFWARM ALGORITHM

For fuzzy weighted association rule mining standard ARM algorithms can be used or at least adopted after some modifications. The Improved Fuzzy Weighted ARM algorithm belongs to the breadth first traversal family of ARM algorithms, developed using tree data structures and works in a fashion similar to the Apriori algorithm [2].

The IFWARM algorithm is given in Table 3.In the Table: C_k is the set of candidate itemsets of cardinality *k*,*w* is the set of weights associated to items *I*. and *L* is the set of frequent item sets.*R* is the final set of generated fuzzy weighted ARS.

TABLE III. IFWARM ALGORITHM

Input: *T*=data set *w*=itemset weights *wsup*=weighted support *wconf*=weighted confidence
Output: *R*=Set of weighted ARs

Main Algorithm:
D=Transform(*T*) //san the database *T* ,and generate a fuzzy database
D using membership function.
L_1 =Initialize(*D*) // initialize parameters, encode attributes values and generate 1-frequent items from the transformed database
For(*k*=2;$L_{k-1} \neq \varnothing$;*k*++)
do begin
C_k =Join(L_{k-1}) //generate *k* candidate set by (*k*-1) frequent itemsets
For each candidates $c \in C_k$
c.supw=Getsupw(*D*,*c*) // calculate weighted support
$L_k = \{c \in C_k \mid c.\text{supw} \geq wsup\}$
$L = \cup_k L_k$;
end
R=Rules(*L*, *wconf*);

Transform (*T*):This step generates a new transformed fuzzy database *D* from the original database by user specified fuzzy sets and weights for each fuzzy set.

Initialize (*D*): The subroutine initializes itemset weights, weighted support and confidence, encode attributes values to number strings to make the same composite items have the same prefix. and generate 1-frequent item sets from the transformed database.

Join (L_{k-1}): This Join step generates C_k from C_{k-1} .before we can prune it using WDCP and check that whether two label are in the same composite items. if two items in a candidate itemsets have the same prefix, we should drop it at first.

Getsupw (*D*, *c*): In this subroutine we calculate weighted support of each candidates items using the formula in session 3. If the fuzzy support value is greater than or equal to the minimum support threshold. we call it a frequent items.

Rules(*L*,*wconf*); The last step calculates weighted confidence of each frequent items using the formula in session 3.if the value greater than or equal to a minimum confidence threshold, we say it is an effective and useful rules.

IV. ALGORITHMS COMPARISON

In this section, we give a comparative analysis of frequent itemset generation between classical

Fuzzy ARM (FARM)[8], MINWAL(W)[5], Normalized Weighted ARM(NWARM)[7],Fuzzy

Weighted ARM (FWARM) [13] and proposed Improved Fuzzy Weighted ARM(IFWARM), In table 4 all the possible itemsets are generated using tables 1 and 2, and the frequent itemsets generated using above algorithms. We denoted by (*X*, *Small*), (*X*, *Large*) , (*Y*, *Small*), (*Y*, *Large*) as A ,B,C,D in turn.

TABLE IV. FREQUENT ITEMSETS COMPARISON

ID	FARM	MINWAL(W)	NWARM	FWARM	IFWARM
1	A(0.675)	A(0.540)	A(0.540)	A(0.540)	A(0.540)
2	B(0.325)	B(0.163)	B(0.163)	B(0.163)	B(0.163)
3	C(0.350)	C(0.210)	C(0.210)	C(0.210)	C(0.210)
4	D(0.650)	D(0.195)	D(0.195)	D(0.195)	D(0.195)
5	AC(0.220)	AC(0.154)	AC(0.152)	AC(0.106)	AC(0.079)
6	AD(0.455)	AD(0.250)	AD(0.223)	AD(0.109)	AD(0.041)
7	BC(0.500)	BC(0.275)	BC(0.274)	BC(0.150)	BC(0.125)
8	BD(0.780)	BD(0.312)	BD(0.302)	BD(0.117)	BD(0.070)

A support threshold for all is set to 2%.Itemsets with a highlighted background indicate frequent itemsets.This experiment is conducted in order to illustrate the effect of item's occurrences and their weights on the generated rules. and we can find that, In column 3 and 4 ,B is not frequent itemset but {BC}and{BD} are frequent.so with weighted settings MINWAL(W),NWARM can not hold Weighted Downward Closure Property.

V. EXPERIMENTAL EVALUATION

To demonstrate the effectiveness of the approach, we performed several experiments using a real customer data set[10],The data is a composite transaction database containing 5900 records and 20 attributes .we select 6 main attributes {country,marital_status,gender,education,age,

year_income} from the data itemsets, country includes three values, and education includes five values, both of them are categorical attributes .we distribute age to three fuzzy regions

{Young,Middle,Old},and year_income to three fuzzy {Low,Middle,High}, marital_status and gender are boolean items, if the item exist ,its fuzzy membership value is 1,or else 0.We conduct the experiment using Matlab on the computer composed of 1GB memory and Intel PE CPU.

In the weighted process, the weight to set the value is of flexibility, primarily on the basis of expert experience ,taking full account of the user's point of view, such as the user will pay more attention to gender on the land the impact of income, you can assign the weight to gender a higher value properties. In this paper, the same attributes are distributed of the same weight, if the user only interest a certain regional attribute, for example, just want to know the income of middle-aged crowd, it can give age of the middle region attribute of the weight to set a higher value, Table 5 shows the attribute value distribution of the weight, and Figure 1 and Figure 2 give membership function of age and year-come.

TABLE V. THE WEIGHT OF ATTRIBUTES

country	marital_status	gender	education	age	Year-income
0.4	0.5	0.6	0.7	0.8	0.8

In this paper, we have used a real dataset in order to demonstrate performance of the proposed approach. We performed two types of experiments based on quality and performance measures. For quality measures, we compared the number of the interesting rules generated using five algorithms described above. In the second experiment, we showed the scalability of these algorithms by comparing the execution time with varying user specified support thresholds.

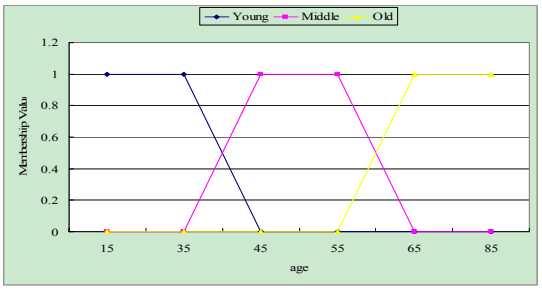

Figure 1. The membership function of age

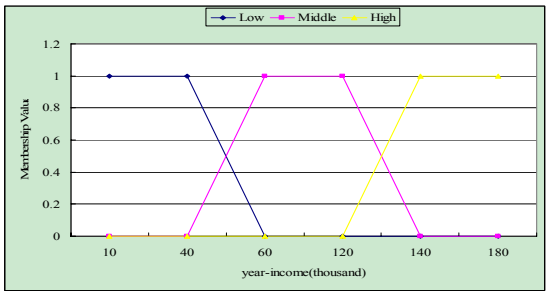

Figure 2. The membership function of year-income

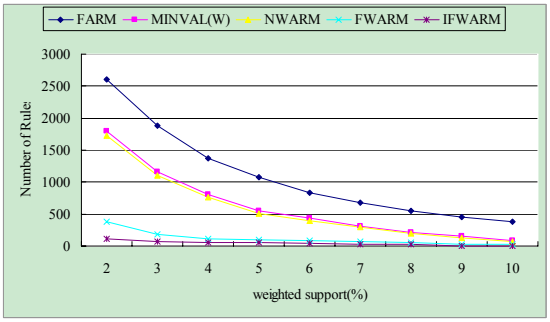

Figure 3. No. of Interesting Rules generated using user specified support(wconf=0.2)

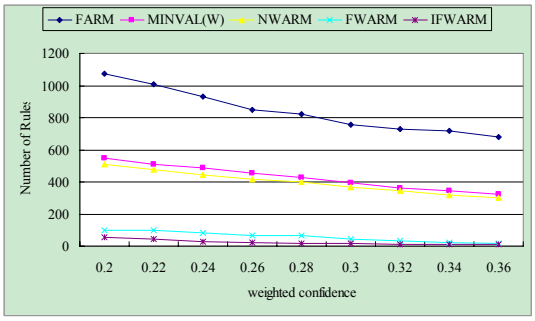

Figure 4. No. of Interesting Rules generated using user specified confidence (wsup=5%)

A. *Quality Measures*

For quality measures, each item is assigned a weight range between [0…1] according to their significance in the dataset. With fuzzy dataset each attribute is divided into three different fuzzy sets as above.

In Figure 3, the x-axis shows confidence thresholds from 2% to 10% and on the y-axis the number of interesting rules .The results show quite similar behavior of the weighted algorithms to classical FARM.As expected the number of rules increases as the minimum support decreases in all cases. Results of proposed IFWARM and FWARM approach are better than MINWAL(W) and NWARM approach, because the formers hold the WDCP and all the potential itemsets are considered from the beginning for pruning using WDCP.the proposed IFWARM is more stable

and effective than FWARM,and the smaller the support value, the more obvious performance. Figure 4 shows the number of interesting rules generated using confidence measures. In all cases, the number of interesting rules is less because the interestingness measure generates fewer rules.

The experiments show that the proposed algorithms produce better results as it uses all the possible itemset and generates rules effectively using valid WDCP.

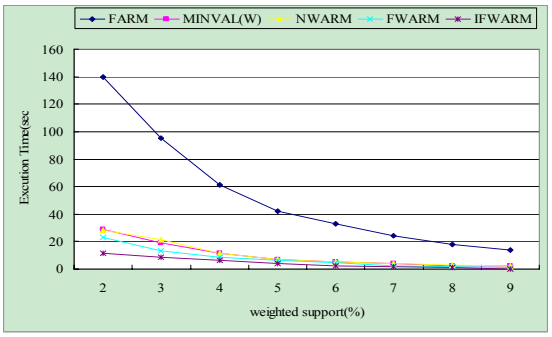

Figure 5. Performance measures: varying weighted support threshold(wconf=0.2)

B. *Performance Measures*

Experiments two compares the execution time of this algorithms. We investigated the effect

On execution time caused by varying the weighted support threshold with fixed data size, in figure 5 a support threshold from 2% to 10% is used again. Due to the way it generates frequent sets i.e. it considers items weights, the weighted algorithms have less execution time than classical FARM. Simulation experiment and the proposed algorithms have less execution time as it uses all the possible itemsets and generates rules effectively using valid WDCP.

VI. CONCLUSION

The paper has presented a generalized approach for mining weighted association rules from databases with quantitative attributes. some classical model of fuzzy association rule mining is adopted to address the issue of invalidation of DCP in weighted association rule mining .and this algorithm can hold WDCP.We have demonstrated the valid WDCP with formal comparisons with some weighted ARM.It is notable that the approach presented here is effective in analyzing databases with fuzzy attributes with weighted settings.

ACKNOWLEDGMENT

In this paper, the research was sponsored by the Science and Technology Plan of Guangzhou City in Guangdong Province (Project No. 2014J4100107)

REFERENCES

[1] R Agrawal, R Srikant. “Fast Algorithms for Mining Association Rules in Large Datebases”[C], *In:Proc 20th Int Conf VLDB*,1994 pp:487-499.

[2] Bodon, F. “A Fast Apriori implementation”, *In ICDM Workshop on Frequent Itemset Mining Implementations*, vol. 90, Melbourne, Florida, USA (2003)

[3] Guoqing Chen, Qiang Wei. “Fuzzy association rules and the extended mining algorithms”. *Information Sciences* 147(2002) pp: 201-228.

[4] Peng Yan. Guoqing Chen, Chris Cornelis, Martine De Cock, Etienne Kerre “Mining positive and negative fuzzy association rules” I*n:Lecture Notes in Computer Science 3213(M.G Negoita, R.J.Howlett, L.C.Jain,eds.)*,Springer-Verlag,2004, pp.270-276

[5] Cai, C.H., Fu, A.W-C., Cheng, C. H., Kwong, W.W. “Mining Association Rules with Weighted Items”. *In: Proceedings of 1998 Intl. Database Engineering and Applications Symposium (IDEAS'98)*, pp 68--77

[6] Delgado, M., et al.: “Fuzzy Association Rules: General Model and Applications”. *IEEE Transactions on Fuzzy Systems* 11(2), pp 214–225 (2003)

[7] Attila Gyenesei. “Minning Weighted Association Rules for Fuzzy Quantitative Items”. *PKDD 2000, LNAI 1910*,pp.416-423 2000.

[8] Didier Dubois.etc “A Systematic Approach to the Assessment of Fuzzy Association Rules” *The 10th International Fuzzy Systems Association World Congress,Istambul*,2003, pp 25-29

[9] Chunqiu Zeng etc MPSQAR: “Minning Quantitative Association Rules Preserving Semantics”, *ADMA 2008,LNAI 5139*,pp.572-570,2008

[10] Mehmet Kaya and Reda Alhajj. “ Effective Mining of Fuzzy Multi-Cross-Level Weighted Association Rules” *ISMIS 2006,LNAI 4203*:pp,399-408

[11] Mehmet Kaya and Reda Alhajj, “Online ming of fuzzy multidimensional weighted association rules”. *Appl Intell(2008)*29: pp 13-14

[12] Shen Bin .etc. “Mining Weighted Generalized Fuzzy Association Rules with Fuzzy Taxonomies.*CIS 2005,Part 1,LNAI 3801*,pp.704-712

[13] M.Sulaiman Khan,Maybin Muyeba and Frans Coenen. “Weighted Association Rule Mining from Binary and Fuzzy Data”. *ICDM 2008,LNAI 5077*,pp.200-212

[14] Guoqing Chen, Qiang Wei “.Fuzzy association rules and the extended mining algorithms”. *Information Sciences* 147(2002) pp 201-228

[15] Hyeong-Joon Kwon and Kwon and Kwang-Seok Hong. “Intelligent Information System Based on a Speech Web Using Fuzzy Association Rule Mining”. *APCHI 2008,LNCS 5068*,pp104-113,2008

[16] Ke Sun and Fengshan Bai. “Mining Weighted Association Rules without Preassigned Weights. ”*IEEE Transactions on knowledge and Data Engineering*,2008(20) pp:489-495