Mining Fuzzy Association Rules from Composite Items

M. Sulaiman Khan[1], Maybin Muyeba[2] and Frans Coenen[3]

Abstract This paper presents an approach for mining fuzzy Association Rules (ARs) relating the properties of composite items, i.e. items that each feature a number of values derived from a common schema. We partition the values associated to properties into fuzzy sets in order to apply fuzzy Association Rule Mining (ARM). This paper describes the process of deriving the fuzzy sets from the properties associated to composite items and a unique Composite Fuzzy Association Rule Mining (CFARM) algorithm founded on the certainty factor interestingness measure to extract fuzzy association rules. The paper demonstrates the potential of composite fuzzy property ARs, and that a more succinct set of property ARs can be produced using the proposed approach than that generated using a non-fuzzy method.

1 Introduction

Association Rule Mining (ARM) is an important and well established data mining topic. The objective of ARM is to identify patterns expressed in the form of Association Rules (ARs) in transaction data sets [1]. The attributes in ARM data sets are usually binary valued but it has been applied to quantitative and categorical (non-binary) data [2]. With the latter, values can be split into ranges such that each range represents a binary valued attribute and ranges linguistically labelled; for example "low", "medium", "high" etc. Values can be assigned to these range attributes using crisp boundaries or fuzzy boundaries. The application

[1] M. Sulaiman Khan (PhD student)
School of computing, Liverpool Hope University, L16 9JD, UK email: khanm@hope.ac.uk
[2] Dr. Maybin Muyeba
School of computing, Liverpool Hope University, L16 9JD, UK email: muyebam@hope.ac.uk
[3] Dr. Frans Coenen
Department of Computer Science, University of Liverpool, L69 3BX email: frans@csc.liv.ac.uk

of ARM using the latter is referred to as fuzzy ARM (FARM) [3]. The objective of fuzzy ARM is then to identify fuzzy ARs. Fuzzy ARM has been shown to produce more expressive ARs than the "crisp" methods [3, 5, 8].

We approach the problem differently in this paper by introducing "Composite Item Fuzzy ARM" (CFARM) whose main objective is the generation of fuzzy ARs associating the "properties" linked with composite attributes [4], i.e., attributes or items composed of sets of sub-attributes or sub-items that conform to a common schema. For example, given an image mining application, we might represent different areas of each image in terms of groups of pixels such that each group is represented by the normalized summation of the RGB values of the pixels in that group. In this case the set of composite attributes (I) is the set of groups, and the set of properties (P) shared by the groups is equivalent to the RGB summation values (i.e. $P = \{R, G, B\}$). Another could be the market basket analysis, where I is a set of groceries, and P is a set of nutritional properties that these groceries posses (i.e. P = {Pr, Fe, Ca, Cu,..}) standing for protein, Iron etc. Note that the actual values (properties) associated with each element of I will be constant, unlike in the case of the image mining example. We note that there are many examples depending on application area but we limit ourselves to these given here.

The term composite item has been used previously in [6, 7] and defined as a combination of several items e.g. if itemset {A, B} and {A, C} are not large then rules {B}→{A} and {C}→{A} will not be generated, but by combining B and C to make a new *composite* item {BC} which may be large, rules such as {BC}→{A} may be generated. In this paper we define composite items differently as indicated earlier, to be an item with properties (see Sect. 3). This definition is consistent in [4] which also defines composite attributes in this manner, i.e. an attribute that comprises two or more sub-attributes.

In this paper, the concept of "Composite item" mining of property ARs is introduced, the potential of using property ARs in many applications and a demonstration of the greater accuracy produced using the certainty factor measure. In addition, it is demonstrated that a more succinct set of property ARs (than that generated using a non-fuzzy method) can be produced using the proposed approach.

The paper is organised as follows; section 2 presents a sequence of basic concepts, section 3 presents the methodology with an example application, Section 4 presents results of the CFARM approach and section 5 concludes the paper with a summary of the contribution of the work and directions for future work.

2. Problem Definition

The problem definition consists of basic concepts to define composite items, fuzzy association rule mining concepts, the normalization process for Fuzzy

Transactions (*FT)* and interestingness measures. Interested readers can see [11] for the formal definitions and more details.

To illustrate the concepts, we apply the methodology using market basket analysis where the set of groceries have a common set of nutritional quantitative properties. Some examples are given in Table 1.

Table 1. Composite items (groceries) with their associated properties (nutrients)

Items/Nutrients	Protein	Fibre	Carbohydrate	Fat	…
Milk	3.1	0.0	4.7	0.2	…
Bread	8.0	3.3	43.7	1.5	…
Biscuit	6.8	4.8	66.3	22.8	…
…	…	…	…	…	…

To illustrate the context of the problem, Table 1 shows composite edible items, with common properties (Protein, Fibre,…). The objective is then to identify consumption patterns linking these properties and so derive fuzzy ARs.

2.1 Basic Concepts

A *Fuzzy Association Rules* [8] is an implication of the form:

$$\text{if } \langle A, X \rangle \text{ then } \langle B, Y \rangle$$

where A and B are disjoint itemsets and X and Y are fuzzy sets. In our case the itemsets are made up of property attributes and the fuzzy sets are identified by linguistic labels.

A *Raw Dataset* D consists of a set of transactions $T = \{t_1, t_2, \cdots, t_n\}$, a set of composite items $I = \{i_1, i_2, \cdots, i_{|I|}\}$ and a set of properties $P = \{p_1, p_2, \cdots, p_m\}$. The "$k^{th}$" property value for the "j^{th}" item in the "i^{th}" transaction is given by $t_i[i_j[v_k]]$. An example is given in Table 2 where each composite item is represented using the notation <label, value>.

The raw dataset D (table 2) is initially transformed into a *Property Dataset* D^p (table 3) which consists of property transactions $T^p = \{t_1^p, t_2^p, \cdots, t_n^p\}$ and a set of property attributes P (instead of a set of composite items I). The value for each property attribute $t_i^p[p_j]$ (the "j^{th}" property in the "i^{th}" property transaction) has a value obtained by aggregating the numeric values for all p_j in t_i (see Table 3). Thus:

$$\text{Prop value}(t_i^p[p_j]) = \frac{\sum_{j=1}^{|t_i|} t_i[i_j[v_k]]}{|t_i|} \quad \textbf{(1)}$$

Table 2 Example raw dataset *D*

TID	Record
1	{<a,{2,4,6}>, <b,{4,5,3}>}
2	{<c,{1,2,5}>, <d,{4,2,3}>}
3	{<a,{2,4,6}>, <c,{1,2,5}>, <d,{4,1,3}>}
4	{<b,{4,5,3}>, <d,{4,2,3}>}

Table 3 Property data set D^p

TID	X	Y	Z
1	3.0	4.5	4.5
2	2.5	2.0	4.0
3	2.3	2.3	4.7
4	4.0	3.5	3.0

Once a property data set D^p is defined, it is then transformed into a *Fuzzy Dataset* D'. A fuzzy dataset D' consists of fuzzy transactions $T' = \{t'_1, t'_2, ..., t'_n\}$ and a set of fuzzy property attributes P' each of which has fuzzy sets with linguistic *labels* $L = \{l_1, l_2, ..., l_{|L|}\}$ (table 4). The values for each property attribute $t_i^p[p_j]$ are *fuzzified* (mapped) into the appropriate membership degree values using a membership function $\mu(t_i^p[p_j], l_k)$ that applies the value of $t_i^p[p_j]$ to a label $l_k \in L$. The complete set of fuzzy property attributes P' is given by $P \times L$.

Composite Itemset Value (CIV) table is a table that allows us to get property values for specific items. The *CIV* table for Table 2 is given in Table 5 below.

Table 4 Properties table

Property	Linguistic values		
	Low	Medium	High
X	$V_k \leq 2.3$	$2.0 \leq V_k \leq 2.3$	$V_k \geq 3.3$
Y	$V_k \leq 3.3$	$3.0 \leq V_k \leq 4.3$	$V_k \geq 4.1$
Z	$V_k \leq 4.0$	$3.6 \leq V_k \leq 5.1$	$V_k \geq 4.7$

Table 5 *CIV* table

Item	Property attributes		
	X	Y	Z
A	2	4	6
B	4	5	3
C	1	2	5
D	4	2	3

Properties Table provides a mapping of property attribute values $t_i^p[p_j]$ to membership values according to the correspondence between the given values to the given linguistic labels. An example is given in Table 5 for the raw data set given in Table 2.

A property attribute set A, where $A \subseteq P \times L$, is a *Fuzzy Frequent Attribute Set* if its fuzzy support value is greater than or equal to a user supplied minimum

support threshold. The significance of fuzzy frequent attribute sets is that fuzzy association rules are generated from the set of discovered frequent attribute sets.

Fuzzy Normalisation is the process of finding the contribution to the fuzzy support value, m', for individual property attributes ($t_i^p[p_j[l_k]]$) such that a partition of unity is guaranteed. This is given by equation 2 where μ is the membership function:

$$t_i'[p_j[l_k]] = \frac{\mu(t_i^p[p_j[l_k]])}{\sum_{x=1}^{|L|} \mu(t_i^p[p_j[l_x]])} \quad (2)$$

If normalisation is not done, the sum of the support contributions of individual fuzzy sets associated with an attribute in a single transaction may no longer be unity which is undesirable.

Frequent fuzzy attribute sets are identified by calculating *Fuzzy Support* values. Fuzzy Support $(Supp_{Fuzzy})$ is calculated as follows:

$$Supp_{Fuzzy}(A) = \frac{\sum_{i=1}^{i=n} \prod_{\forall[i[l]]\in A} t_i'[i[l]]}{n} \quad (3)$$

where $A = \{a_1, a_2, ..., a_{|A|}\}$ is a set of property attribute-fuzzy set (label) pairs. A record t_i' "satisfies" A if $A \subseteq t_i'$. The individual vote per record, t_i is obtaining by multiplying the membership degree associated with each attribute-fuzzy set pair $[i[l]] \in A$.

2.2 Interestingness Measures

Frequent attribute sets with fuzzy support above the specified threshold are used to generate all possible rules. *Fuzzy Confidence* $(Conf_{Fuzzy})$ is calculated in the same manner that confidence is calculated in classical ARM:

$$Conf_{Fuzzy}(A \rightarrow B) = \frac{Supp_{Fuzzy}(A \cup B)}{Supp_{Fuzzy}(A)} \quad (4)$$

The Fuzzy Confidence measure $(Conf_{Fuzzy})$ described does not use $Supp_{Fuzzy}(B)$ but the *Certainty* measure $(Cert)$ addresses this. The certainty measure is a statistical measure founded on the concepts of *covariance (Cov)* and *variance (Var)* and is calculated as follows:

$$Cert(A \rightarrow B) = \frac{Cov(A,B)}{\sqrt{Var(A) \times Vat(B)}} \quad (5)$$

The value of certainty ranges from -1 to +1. We are only interested in rules that have a certainty value greater than 0. As the certainty value increases from 0 to 1, the more related the attributes are and consequently the more interesting the rule.

3. Methodology

To evaluate the approach, a market basket analysis data set with 600 composite edible items is used and the objective is to determine consumers' consumption patterns for different nutrients using RDA. The properties for each item comprised the 27 nutrients contained in the government sponsored RDA table (a partial list consists of Biotin, Calcium, Carbohydrate, ..., Vitamin K, Zinc). These RDA values represent a CIV table used in the evaluation. The property data set will therefore comprise $600 \times 27 = 16200$ attributes. The linguistic label set L was defined as follows L – {Very Low (VL), Low (L), Ideal (I), High (H), Very High (VH)}. Thus the set of fuzzy attributes $A = P \times L$ has $27 \times 5 = 135$ attributes. A fragment of this data (properties table) is given in Table 6.

Table 6 Fragment of market basket properties table[4].

Nutrients / Fuzzy Ranges	Very Low				Low				Ideal				High				Very High		
	Min	Core		Max	Min	Core		Max	Min	Core		Max	Min	Core		Max	Min	Core	
Fiber	0	1	10	15	10	15	20	25	20	25	30	35	30	33	38	39	35	40	…
Iron	0	6	8	12	8	12	16	18	16	18	19	20	19	20	22	23	22	23	…
Protein	0	1	15	30	10	20	35	40	35	40	60	65	60	65	75	80	75	70	…
Vitamin	0	15	150	200	150	200	300	400	300	350	440	500	440	490	550	600	550	600	…
Zinc	0	0.8	8	10	8	10	15	20	15	20	30	40	30	40	46	50	46	50	…
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

A representative fragment of a raw data set (T), comprising edible items, is given in Table 7(a). This raw data is then cast into a properties data set (T^P) using the given CIV/RDA table to give the properties data set in Table 7(b). It is feasible to have alternative solutions here but we choose to code fuzzy sets {very

[4] Values could be in grams, milligrams, micrograms, International unit or any unit. Here Min is the minimum value i.e. α, Core is the core region β, δ and Max is the maximum value γ in the trapezoidal fuzzy membership function.

low, low, ideal, high, very high} with numbers {1, 2, 3, 4, 5} for the first nutrient (Pr), {6, 7, 8, 9, 10} for the second nutrient (Fe) etc [9]. Thus, data in Table 7(c) can be used by any binary ARM algorithm.

Table 7 (a) [a]

TID	Items
1	X, Z
2	Z
3	X,Y, Z
4	…

[a] Raw data (T)

Table 7 (b) [b]

TID	Pr	Fe	Ca	Cu
1	45	150	86	28
2	9	0	47	1.5
3	54	150	133	29.5
4	…	…	…	…

[b] Property data set (T^P)

Table 7 (c) [c]

TID	Pr	Fe	Ca	Cu
1	3	8	13	16
2	1	6	12	16
3	3	8	15	16
4	…	…	…	...

[c] Conventional ARM data set

This approach only gives us, the total support of various fuzzy sets per nutrient and not the degree of (fuzzy) support. This directly affects the number and quality of rules as stated in Sect. 4. To resolve the problem, the fuzzy approach here converts RDA property data set, Table 7(b), to linguistic values (Table 8) for each nutrient and their corresponding degrees of membership reflected in each transaction.

Table 8 Linguistic transaction file

TID	Protein (Pr)					Iron (Fe)					
	VL	L	Ideal	H	VH	VL	L	Ideal	H	VH	…
1	0.0	0.7	0.3	0.0	0.0	0.0	0.0	0.8	0.2	0.0	…
2	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	…
3	0.0	0.0	0.9	0.1	0.0	0.0	0.0	0.8	0.2	0.0	…
4	…	…	…	…	…	…	…	…	…	…	…

Table 8 shows only two nutrients, Pr and Fe (i.e. a total of 10 fuzzy sets).

4. Experimental Results

To demonstrate the effectiveness of the approach, we performed several experiments using a real retail data set [10]. The data is a transactional database containing 88,163 records and 16,470 unique items. For the purpose of the experiments we mapped the 600 item numbers onto 600 products in a real RDA table. Results in [11] were produced using synthetic dataset. In this paper, an improvement from [11] is that we have used real dataset in order to demonstrate the real performance of the proposed approach and algorithm.

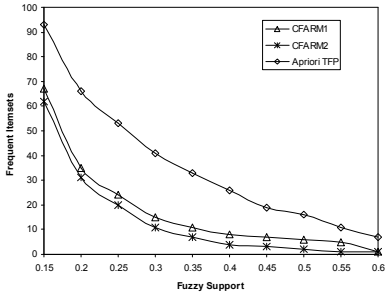
The Composite Fuzzy ARM (CFARM) algorithm is a *breadth first traversal* ARM algorithm, uses tree data structures and is similar to the Apriori algorithm

[1]. The CFARM algorithm consists of several steps. For more details on algorithm and pseudo code please see [11].

4.1 Quality Measures

In this section, we compare Composite Fuzzy Association Rule Mining (CFARM) approach against standard Quantitative ARM (discrete method) with and without normalisation. We compare the number of frequent sets and the number of rules generated using both the confidence and the certainty interestingness measure. Fig. 1 demonstrates the difference between the numbers of frequent itemsets generated using Quantitative ARM approach with discrete intervals and CFARM with fuzzy partitions. CFARM1 uses data without normalisation and CFARM2 uses normalised data. For standard Quantitative ARM, we used Apriori-TFP algorithm [12]. As expected the number of frequent itemsets increases as the minimum support decreases.

Figure 1 Number of frequent Itemsets generated using fuzzy support measures



It is clear from the results that the algorithm that uses discrete intervals produces more frequent itemsets than fuzzy partitioning method. This is because standard ARM (using discrete intervals) generates numerous artificial patterns resulting from the use of crisp boundaries. Conversley, fuzzy partitioning methods generate more accurately the true patterns in the data set due to the fact that it considers actual contribution of attributes in different intervals. CFARM2 produces comparatively less frequent itemsets than CFARM1, because the average contribution to support counts per transaction is greater without using normalization than with normalization.

Fig. 2 shows the number of rules generated using user specified fuzzy confidence. Fig. 3 shows the number of interesting rules generated using certainty measures values. Certainty measures (Fig. 2) generate fewer, but arguably better, rules than the confidence measure (Fig. 2). In both cases, CFARM2 generates less rules as compared to CFARM1; this is a direct consequence of the fact that CFARM 2 generates fewer frequent itemsets due to using normalised data.

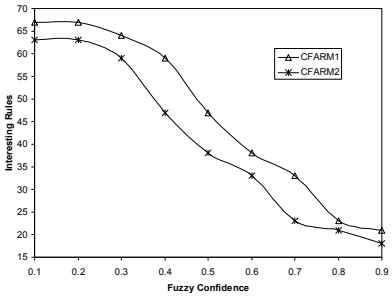


Figure 2 Interesting Rules using confidence

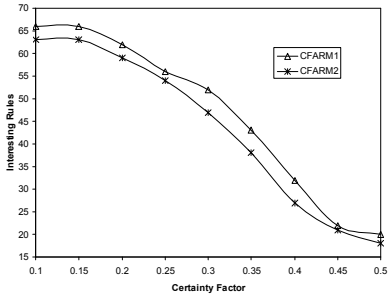


Figure 3 Interesting Rules using certainty

In addition, the novelty of the approach is its ability to analyse datasets comprised of composite items where each item has a number of property values such as the nutritional property values used in the application described here.

4.2 Performance Measures

For performance measures, we investigated the effect on algorithm execution time caused by varying the number of attributes and the data size with and without normalization using a support 0.3, confidence 0.5 and certainty 0.25. The dataset was partitioned into 9 equal partitions labelled 10K, 20K, …, 90K to obtain different data sizes We used all 27 nutrients.

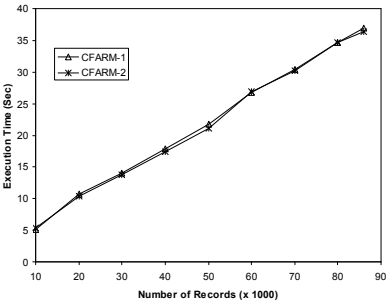


Figure 4 Execution time: No. of Records

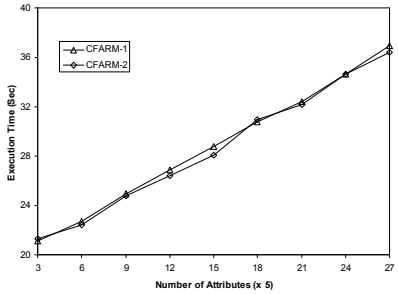


Figure 5 Execution time: No. of Attributes

Fig. 4 shows the effect on execution time by increasing the number of records. From Fig. 4 it can be seen that both algorithms have similar timings while the execution time increasing with the number of records. Fig. 5 shows the effect on execution time by varying numbers of attributes. Each property attribute has 5 fuzzy sets associated to it, therefore using 27 attributes, we have 135 columns.

However the experiments also show that the CFARM algorithm scales linearly with the number of records and attributes.

5. Conclusion and Future Work

A novel approach was presented for extracting fuzzy association rules from so-called composite items where such items have properties defined as quantitative (sub) itemsets. The properties are then transformed into fuzzy sets. The CFARM algorithm produces a more succinct set of fuzzy association rules using fuzzy measures and certainty as the interestingness measure and thus presents a new way for extracting association rules from items with properties. This is different from normal quantitative ARM. We also showed the experimental results with market basket data where edible items were used with nutritional content as properties. Of note is the significant potential to apply CFARM to other applications where items could have composite attributes even with varying fuzzy sets between attributes. We have shown that we can analyse databases with composite items using a fuzzy ARM approach.

References

1. Rakesh Agrawal, Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules in Large Databases, VLDB, (1994) 487-499
2. R. Srikant and R. Agrawal, Mining Quantitative Association Rules in Large Relational Tables, Proc. ACM SIGMOD Conf. on Management of Data. ACM Press, (1996) 1 - 12
3. Kuok, C., Fu, A., Wong, H.: Mining Fuzzy Association Rules in Databases, ACM SIGMOD Record Vol. 27, (1), (1998) 41-46
4. W. Kim, E. Bertino and J. Garza, Composite objects revisited, ACM SIGMOD Record, Vol. 18, (2) (1989) 337-347
5. Dubois, D. E. Hüllermeier, H. Prade, A Systematic Approach to the Assessment of Fuzzy Association Rules, DM and Knowledge Discovery Journal, Vol. 13(2), (2006) 167-192
6. X. Ye and J. A. Keane, Mining Composite Items in Association Rules, Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, (1997) 1367-1372
7. K. Wang, J. K. Liu and W. Ma, Mining the Most Reliable Association Rules with Composite Items, in Proc. ICDMW'06, (2006), 749-754
8. M. Delgado, N. Marin, D. Sanchez and M. A. Vila, Fuzzy Association Rules, General Model and Applications, IEEE Transactions on Fuzzy Systems, 11(2) (2003) 214–225
9. M. Sulaiman Khan, M. Muyeba, F. Coenen, On Extraction of Nutritional Patterns (NPS) Using Fuzzy Association Rule Mining, proc. Intl. Conf. on Health Informatics, Madeira, Portugal, (2008), Vol. 1, 34 - 42.
10. Brijs T. and et al., The use of association rules for product assortment decisions: a case study, proc. 5th Intl. Conf. on Knowledge Discovery and Data Mining (1999) 254-260.
11. M. Sulaiman Khan, M. Muyeba, F. Coenen, A Framework for Mining Fuzzy Association Rules from Composite Items, to appear in ALSIP (PAKDD) 2008, Osaka, Japan.
12. Coenen, F., Leng, P., Goulbourne, G.: Tree Structures for Mining Association Rules, Data Mining and Knowledge Discovery, Vol. 8, No. 1, (2004) 25 – 51