An effective Fuzzy Healthy Association Rule Mining Algorithm (FHARM)

M. Sulaiman Khan, Maybin Muyeba, Christos Tjortjis

Abstract--- **This paper presents an effective Healthy Association Rule Mining (HARM) algorithm by introducing new quality measures to generate more interesting rules using itemset nutrient information. Our previous method for analyzing healthy buying patterns from quantitative attributes (or nutrient information) by interval partitions using the classical *Apriori* algorithm was unable to generate quality rules. This was partly because using basic itemset support is not an appropriate approach as it only gives the total support of various fuzzy sets per nutrient and not the degree of support. In this paper we propose an effective and efficient new Fuzzy Healthy Association Rule Mining Algorithm (FHARM) that produces more interesting and quality rules. In this approach, edible attributes are filtered from transactional input data by projections and are then converted to Required Daily Allowance (RDA) numeric values. The average RDA database is then converted to a fuzzy database that contains normalized fuzzy attributes comprising different fuzzy sets. Analysis of nutritional information is then performed from the converted normalized fuzzy transactional database. The paper presents various performance tests and interestingness measures to demonstrate the effectiveness of the approach and proposes further work on evaluating our approach with other generic fuzzy association rule algorithms.**

I. INTRODUCTION

Data mining is an important research area and Association Rule Mining (ARM) is one of the most investigated and well established data mining techniques to determine customer buying patterns or commonly known as association rules (ARs) [2, 12, 13, 18]. Most of the techniques for extracting valid, authentic and interesting patterns are quantitative in nature [6, 19]. Qualitative attributes are also required in order to fully exploit object attributes present in the data set. Fuzzy approaches are used to extract interesting rules from qualitative or linguistic terms in relational and transactional databases [4, 5, 14].

The term Healthy Buying Patterns (HBP) was introduced in [17] and signifies the level of nutritional content in an association rule per item. In our previous work we analyzed healthy buying patterns from quantitative attributes by their interval partitions.

We used an efficient tree-based ARM algorithm [12] to generate healthy buying patterns, but were unable to generate quality rules using user specified confidence and support. This is partly due to the partitioning of fuzzy sets and their conversion into Boolean values. This produces the total support of various fuzzy sets per nutrient but not the actual degree of support.

In this paper we propose a fuzzy algorithm for healthy association rule mining, where a transactional database is converted into a database that contains the average RDA of nutrient values per item. This database is then converted into a fuzzy database with fuzzy attributes, according to the nutrients intake. The fuzzy database contains the actual fuzzy membership degrees of fuzzy sets for each particular item (e.g. values 0.1, 0.3, 0.5, 0.1, 0.0 for fuzzy attributes very low, low, ideal, high and very high respectively). We show the effectiveness of this new method by applying it on different datasets. Our contributions are that edible attributes are used in our algorithms with an RDA table, a fuzzy normalization process and correlation analysis produce effective rules and records good performance.

The remainder of the paper is organised as follows: section 2 presents background and related work; section 3 gives a problem definition; section 4 discusses the methodology; section 5 details the proposed algorithm; section 6 reviews experimental results, and section 7 concludes the paper with directions for future work.

II. BACKGROUND AND RELATED WORK

Many ARM algorithms are more concerned with efficient implementations than producing effective rules [2, 3, 13, 14, 16]. Again, in almost all ARM algorithms, thresholds (both confidence and support) are crisp values. This support specification may not suffice for queries and rule representations that require generating rules that have linguistic terms such as “low protein” etc. Fuzzy approaches [4, 5, 7, 17] deal with quantitative attributes [6] by mapping numeric values to Boolean values. Detailed overviews for fuzzy association rules are given in [10, 15]. Mining nutrient associations among itemsets is a new type of ARM algorithm which attempts to investigate HBP by analysing nutrition consumption patterns [17]. In [8], fuzzy associations are presented, where a reduced table is used to effectively minimise the complexity of mining such rules. The authors also present mining for nutrients in the antecedent part of the rule, but it is not clear how the fuzzy nutrient values are aggregated and largely, how membership functions are used. Our algorithm’s ultimate goal is to

M. Sulaiman Khan and Maybin Muyeba are with the School of Computing, Liverpool Hope University, Liverpool, L16 9JD, UK (email: khanm@hope.ac.uk, muyebam@hope.ac.uk)

Christos Tjortjis is with the School of Computer Science, University of Manchester, Manchester, PO Box 88, M60 1QD, UK (email: christos.tjortjis@manchester.ac.uk).

determine customers' buying patterns for healthy foods, which can easily be evaluated using RDA standard tables. Other related work deals with building a classifier using fuzzy ARs in biomedical applications [9].

III. Problem Definition

A major problem in discretising quantitative attributes using interval partitions, discussed in [10], is that of sharp boundary problems where support thresholds leave out transactions on the boundaries of these intervals. Thus the approach to resolve this, using fuzzy sets, is adopted in this paper. Fuzzy Association Rule Mining is the problem of discovering frequent itemsets using fuzzy sets in order to handle the quantitative attributes in transactional and relational databases.

In this section, first we will describe the concept of fuzzy association rule mining and the fuzzy approach we have adopted suitable for our HARM problem. Normalization process for Fuzzy Transactions (*FT)* and rules interestingness measures will also be discussed later in this section.

A. Fuzzy Association Rules

For a given database D with transactions $T = \{t_1, t_2, t_3, ..., t_n\}$ and converted fuzzy transactions $FT = \{ft_1, ft_2, ft_3, ..., f_{t_n}\}$ with attributes $I = \{i_1, i_2, i_3, ..., i_n\}$ and the fuzzy sets $F = \{f_{i_1}, f_{i_2}, ..., f_{i_n}\}$ associated with each attribute in *I*.

Table 1. Set of ordinary transactions

D	i_1	i_2	i_3
t_1	0	1	1
t_2	0	0	1
t_3	1	1	1
t_4	1	0	0
t_5	0	1	0

Table 2. Set of edible fuzzy transactions

E	$fv_1(i_1)$	$fv_2(i_1)$	$fv_3(i_1)$	$fv_4(i_1)$	$fv_5(i_1)$	$fv_1(i_3)$	$fv_2(i_3)$	$fv_3(i_3)$	$fv_4(i_3)$	$fv_5(i_3)$
f_{t1}	.5	.3	.2	0	0	.4	.4	.2	0	0
f_{t2}	0	0	.7	.2	.1	0	0	.2	.3	.5
f_{t3}	1	0	0	0	0	.1	.9	0	0	0
f_{t4}	0	0	.6	.2	.2	0	0	.8	.1	.1
f_{t5}	.8	.1	.1	0	0	.7	.1	.2	0	0

A fuzzy transaction is a special case of transformed ordinary transaction (table 1) and nonempty fuzzy subset of *I where* $T \subseteq I$ .In table 2 an item i_k and transaction t_j contains a value (membership degree) in [0, 1]. The membership degree of i_k in t_j is $t_j(i_k)$. Without loss of generality, we also define edible set of items $E \subseteq I$ where any $i_j \in E$ consists of quantitative nutritional information $\bigcup_{k=1}^{p} i_j^{\,k}$, where each $i_j^{\,k}$ is given as standard RDA numerical ranges and consists of *p* nutrients. Each quantitative item i_j is divided into various fuzzy sets $f(i_j)$ and $m_{i_j}(l, v)$ denotes the membership degree of v in the fuzzy set l , $0 \le m_{i_j}(l, v) \le 1$ as shown in table 2.

A fuzzy quantitative rule represents each item as (item, value) pair. Fuzzy association rules are thus expressed in the following form:

If X is A satisfies Y is B

For example (Protein is high) => (fats is ideal). In the above rule, $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$ are itemsets, where $X \subset I, Y \subset I$, and $X \cap Y = \varnothing$. Sets $A = \{f_{x1}, f_{x2}, ..., f_{xn}\}$ and $B = \{f_{y1}, f_{y2}, ..., f_{yn}\}$ contain the fuzzy sets associated with the corresponding attributes in X and Y, for example (protein, low), (protein, ideal), (protein, high). The semantics of the rule is that when 'X is A' is satisfied, we can imply that 'Y is B' is also satisfied, which means there are sufficient records that contribute their votes to the attribute fuzzy set pairs and the sum of these votes is greater than the user specified threshold which could be crisp or fuzzy.

B. Fuzzy Transactions Normalization Process

As mentioned above each quantitative item i_j in t_k is divided into various fuzzy sets $f_{(i_j)}$ and $m_{i_j}(l, v)$ denotes the membership degree of v in the fuzzy set l $0 \le m_{i_j}(l, v) \le 1$. For each fuzzy transaction $t \in E$ (edible items), a normalization process to find significance of an items contribution to the degree of support of a transaction in order to guarantee a partition of unity is given by the equation (1):

$$m'_{i_j} = \frac{m_{i_j}(l, t.i_j)}{\sum_{l=1}^{f(i_j)} m_{i_j}(l, t.i_j)} \qquad \textbf{(1)}$$

Without normalisation, support of an individual fuzzy item could increase in a transaction. The normalisation process ensures fuzzy membership values for each nutrient are consistent and are not affected by boundary values.

C. Fuzzy Support and Confidence

The problem of mining fuzzy association rules is given following a similar formulation in [15]. To generate Fuzzy

Support (FS) value of an item set X with fuzzy set A, we use the equation (2):

$$FS(X,A)=\frac{\sum_{t_i\in T}\prod_{x_j\in X} m_{a_j}\in A(t_i[x_j])}{|E|} \qquad (2)$$

A quantitative rule represents each item as <item, value> pair. In the above equation we have used arithmetic mean averaging operator for fuzzy nutrients aggregation of candidate itemsets in a transactional database and used mul operator for fuzzy union of candidate items in a transaction. min or max operators can also be used but mul provides us the simplest and reasonable results as shown in table 3. In case when the fuzzy transactions are not normalized mul is more suitable because it takes the degrees of all items in a transaction into account.

Table 3. Effect of fuzzy *mul* operator

i_1	i_2	i_3	i_4		*Max*	*Min*	*Mul*
.2	.6	.7	.9	→	.9	.2	.075
.9	.8	.5	.6	→	.9	.5	.216
.7	0	.75	.8	→	.8	0	0
.3	.9	.7	.2	→	.9	.2	.037

For a rule $<X,A>\rightarrow<Y,B>$, the fuzzy confidence value (FC) where $X\cup Y=Z, A\cup B=C$ is given by equation (3):

$$FC(<X,A>\rightarrow<Y,B>)=\frac{\sum_{t_i\in T}\prod_{z_j\in X} m_{c_j}\in C(t_i[z_j])}{\sum_{t_i\in T}\prod_{x_j\in X} m_{a_j}\in A(t_i[x_j])} \qquad (3)$$

where each $z\in\{X\cup Y\}$. For our approach, $X,Y\subset E$, where E is a projection of edible items from D. Depending on the query, each item i_j specified in the query and belonging to a particular transaction, is split or converted into p nutrient parts $\bigcup_{k=1}^{p} i_j^{\,k}, 1\le j\le m$. For each transaction t, the bought items contribute to an overall nutrient k by averaging the total values of contributing items i.e. if items i_3, i_4 and i_7 are in a transaction t_1 and all contain nutrient k=5 in any proportions, their contribution to nutrient 5 is $\sum\frac{|i_j^{\,5}|}{3}$, j ∈ {3,4,7}. These values are then aggregated into an RDA table with a schema of nutrients (see table 5, section 4) and corresponding transactions. We use the same notation for an item i_j with nutrient k, $i_j^{\,k}$ as item or nutrient i_k in the RDA table. Given that items i_k are quantitative (fuzzy) and we need to find fuzzy support and fuzzy confidence as defined, we introduce membership functions for each nutrient or item since for a normal diet intake, ideal intakes for each nutrient vary. However, five (5) fuzzy sets for each item are defined as {very low, low, ideal, high, very high} based on expert analysis on nutrition.

Based on this analysis, examples of fuzzy membership functions for nutrient Protein is shown in figure 1. The functions assume a trapezoidal shape since nutrient values in excess or in deficiency mean less than ideal intake according to expert knowledge. Ideal nutrients can assume value 1 naturally, but this value could be evaluated computationally to 0.8, 0.9 in practical terms.

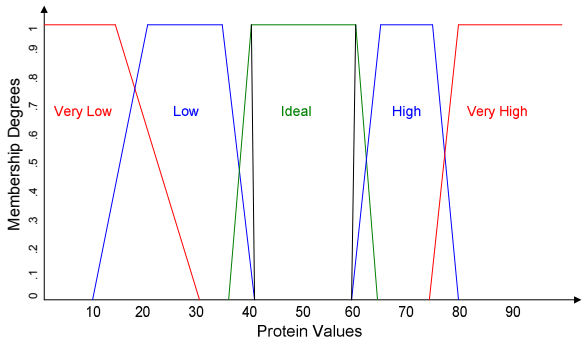

Figure 1: Fuzzy membership functions

$$\mu(x,\alpha,\beta,\gamma,\delta)=\begin{cases}1, & x\in[\beta,\gamma]\\ \frac{(x-\alpha)}{(\beta-\alpha)}, & x\in[\alpha,\beta]\\ \frac{(\delta-x)}{(\delta-\gamma)}, & x\in[\gamma,\delta]\\ 0, & otherwise\end{cases} \qquad (4)$$

Equation 4 [11] represents all nutrient membership degrees of a nutrient value "x". The input nutrient value x has "ideal" values ranging from β to γ, with lowest value α and highest value δ. The task is to determine a membership value for x from equation 4.

Note that equation 4 gives values equal to $m_{i_k}(l,v)$ in equations 1, 2 and 3. We can then handle any query after a series of data transformations and fuzzy function evaluations of associations between nutritional values. For missing nutrient values or so called "trace" elements, the fuzzy function evaluated zero degree membership.

D. Interestingness Measures

Measures of interestingness other than standard support and confidence are required in order to evaluate the quality of fuzzy association rules. The quality measure for a rule to be interesting is called certainty factor [15]. A rule can be considered interesting if the fuzzy set union of antecedent and the consequent has enough significance and the rule has adequate certainty. A measure of significance for a rule is similar to equation (3) and we have adopted it as the confidence of a rule. The certainty factor is determined by computing the fuzzy correlation of antecedent and the consequent of the rule. We have used Pearson's product-moment correlation coefficient between attributes which is

different from the general statistical usage of correlation because in association rule mining $X \Rightarrow Y \neq Y \Rightarrow X$.

The correlation $Corr(X,Y)$ between two variables *X* and *Y* with expected values *E(X)* and *E(Y)* and standard deviations σx and σy is defined as:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma x \sigma y}$$

$$Corr(X,Y) = \frac{E(X.Y) - E(X).E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}}$$

where *E* is the expected value of the variables and *cov* is covariance. We can transform the above correlation equation to find the certainty factor between two or more fuzzy attributes and can calculate fuzzy correlation as:

$$Corr_{Fuzzy}(<X,A>,<Y,B>) = \frac{Cov(<X,A>,<Y,B>)}{\sqrt{Var(X,A)}\sqrt{Var(Y,B)}} \quad (5)$$

where $Cov(<X,A><Y,B>)$ is the fuzzy covariance between the i_j value in *X* and the i_k value in *Y*, and can be determine as:

$$Cov(<X,A><Y,B>) = E(Z,C) - E(X,A) \times E(Y,B)$$

and fuzzy variance of X and Y is obtained as:

$$Var(X,A) = E[(X,A)^2] - E[(X,A)]^2$$

$$Var(Y,B) = E[(Y,B)^2] - E[(Y,B)]^2$$

$$E[(X,A)] = \frac{\sum_{t_i \in T} \prod_{x_j \in X} m_{a_j} \in A(t_i[x_j])}{|E|}$$

$$E[(Y,B)] = \frac{\sum_{t_i \in T} \prod_{y_j \in Y} m_{b_j} \in B(t_i[y_j])}{|E|}$$

The value of correlation ranges from -1 to +1. Value -1 means no correlation and +1 means maximum correlation. In our problem, only positive values can be considered as the degree of relation. As the certainty value increases from 0 to 1, the more related the attributes are and consequently the more interesting they are. Therefore if the rule "IF *Protein* is *low* THEN *Vitamin A* is *high*" holds, then the certainty value should be at least greater than zero. This could mean customers prefer to buy more vitamin related items to protein ones and the HBP value is simply the certainty value obtained (see section 6.1)

IV. METHODOLOGY

We have developed an algorithm called Fuzzy Healthy Association Rule Mining algorithm (FHARM). FHARM can deal with other kinds of transactional and relational databases to generate fuzzy association rules using quantitative attributes.

In our method, input data from the transactional file (as seen in table 4), are projected on-the-fly into a database of edible items, thereby reducing the number of items in the transactions and possibly transactions too. The latter occurs because some items may be non-edible and are not needed for nutrition evaluation. This new input data is converted into an RDA transaction table (table 5) with each edible item expressed as a quantitative attribute and then aggregating all such items per transaction.

At this point, two solutions may exist for the next mining step. One is to code fuzzy sets {very low, low, ideal, high, very high} as, for example,{1, 2, 3, 4, 5}, for the first item or nutrient, {6, 7, 8, 9, 10} for the second nutrient and so on [17]. The encoded data (table 6) can be mined by any non-binary type association rule algorithm to find frequent item sets and hence association rules. This approach only gives us, for instance, the total support of various fuzzy sets per nutrient and not the degree of support as expressed in equations 1 and 2. This directly affects the number and quality of rules as proved in section 6. To overcome this, a fuzzy approach has been adopted. In our fuzzy approach we convert RDA transactions (table 5) to linguistic values (table 7) for each nutrient and corresponding degrees of membership for the fuzzy sets they represent above or equal to a fuzzy support threshold. Each transaction then will have fuzzy values {very low, low, ideal, high, very high} for each nutrient present in every item of that transaction.

Table 4. Transaction file

TID	Items
1	X, Z
2	Z
3	X,Y, Z
4	…

Table 5. RDA transactions

TID	Pr	Fe	Ca	Cu
1	20	10	30	60
2	57	31	70	2
3	99	0	64	80
4	…	…	…	…

Table 6. Fuzzy transactions

TID	Pr	Fe	Ca	Cu
1	1	7	15	24
2	3	10	11	20
3	5	6	15	25
4	…	…	…	...

Table 7. Linguistic transaction file

TID	VL	L	Ideal	H	VH	VL	L	Ideal	H	VH	…
1	0.03	0.05	0.9	0.01	0.01	0.2	0.1	0.8	0	0.7	…
2	0.2	0.1	0.0	0.7	0.1	0.23	0.2	0	0.5	0.1	…
3	0.7	0.2	0.03	0.15	0.12	0	0.5	0.3	0.3	0.11	…
4	…	…	…	…	…	…	…	…	…	…	…

Table 7 shows only two nutrients (i.e. a total of 10 fuzzy sets). A tree data structure is then used to store frequent itemsets using these values (linguistic value and membership degree) and large itemsets found based on the fuzzy support threshold. To obtain the degree of fuzzy support, we use equations 1 and 2 on each fuzzy set for each nutrient and then apply measures of interestingness and quality (equation

5). We then get Association Rules with corresponding HBP values.

V. FHARM ALGORITHM

For fuzzy association rule mining standard ARM algorithms can be used or at least adopted after some modifications [17]. Less attention has been given to developing efficient algorithms for fuzzy association rule mining [7] but still, there are some contributions in this area [10, 15, 16].

An efficient algorithm was needed for the FHARM methodology because a lot of pre-processing (filtration, conversions, normalization) and mining steps are involved in the generation of healthy buying patterns (HBPs). In our problem, ordinary Boolean ARM algorithms are inappropriate. Also the conversion process from ordinary transactions to average RDA transactions and then RDA transactions to Fuzzy Transactions is quite different from other fuzzy attribute extensions of ARM. Careful attention is needed in attribute partitioning because we do not employ any clustering technique; we do this manually from input data with nutritional values given an RDA table.

The *Fuzzy HARM Algorithm* belongs to the *breadth first traversal* family of ARM algorithms, developed using tree data structures [12] and it works in a fashion quite similar to the Apriori algorithm [2]. Also, our implementation approach is different from [10, 15] by avoiding an extra database scan to find correlation values, thus increasing efficiency. FHARM algorithm consists of four major steps:

1. Filtration and transformation of ordinary transactional database into a database with edible average RDA transactions.
2. Appropriate and accurate transformation of RDA transactions into a database containing fuzzy extensions. Normalization of this database.
3. Candidate generation and search for all fuzzy frequent itemsets within candidates that have fuzzy support higher than user specified minimum support.
4. Use of frequent itemsets to generate the desired Healthy Buying Patterns of the form [Protein intake = Ideal → Carbohydrate intake = Low] by calculating the fuzzy confidence and correlation values.

Algorithm Notations:

$I_{edibles}$	Edible items
I_{rda}	Converted edible attributes
D	transactional database
T	set of ordinary transactions
RDA	real nutritional standard RDA table
$T_{edibles}$	transactions with edible items
T_{rda}	RDA converted transactions
D_{rda}	RDA transactions database
T_{fuzzy}	fuzzy transactions
D_{fuzzy}	fuzzy transactions database
F_k	set of frequent k-itemsets
C_k	set of candidates k-itemsets
I	set of complete item sets
$mincorr$	minimum correlation value
$minsupp$	minimum support
$minconf$	minimum confidence

FHARM comprises of the following algorithmic components:

RDAConverter(T, RDA, $I_{edibles}$)

1. $\forall$ T in D
2. $\forall$ I in T
3. if (edible = = check(I_j, $I_{edibles}$)
4. $T_{edible} = T_{edible} \bigcup (I_j)$
5. T_{rda} = averageRDA(T_{edible})
6. D_{rda}.= *wrirte*(T_{rda})
7. end;

RDA-FuzzyConverter(T_{rda}, *FuzzyNutrients*)

1. $\forall$ T_{rda} in D_{rda}
2. $\forall$ I_{rda} in T_{rda}
3. fuzzyattr=getFuzzyAttr(I_{rda},*FuzzyNutrients*)
4. $T_{fuzzy} = T_{fuzzy} \bigcup (fuzzyattr)$
5. D_{fuzzy} = *wrirte*(T_{fuzzy})
6. end;

FHARM(minsupp, minconf, mincorr, T_{fuzzy})

1. k = 0; $C_k = \varnothing$; $F_k = \varnothing$
2. do
3. k = k+1
4. if(k = = 1)
5. C_k = GenerateFirstCandidates (T_{Fuzzy})
6. else
7. C_k = Generatecandidates(F_{k-1})
8. $\forall$ C_k
9. count = CountSupport(C_k)
10. C_k = PruneCandidates(C_k, *count*, *minsupp*)
11. C_k = CalcSignificance(C_k, *minconf*)
12. F_k = GenerateFrequentItemsets(C_k,*minconf*)
13. $F = F \bigcup F_k$
14. while(C_k .count> k)
15. cfactor = CalcCertainty(F, *mincorr*)
16. *Output (* Rules(F , *mincorr, cfactor*))

VI. EXPERIMENTAL RESULTS

In this paper we enhance our previous work [17]. To show the quality, performance and effectiveness of our new approach, we performed several experiments using T10I4D100K dataset containing simulated market basket data [generated by the IBM Almaden Quest research group][1]. The data contains 100K transactions and 1000 items. We considered 600 edible items out of the 1000 and used a real nutritional standard RDA table to derive fuzzy sets.

A. Experiment One: (Quality Measures)

This experiment shows how the new fuzzy HARM approach gives more interesting rules than the previous one using Apriori-TFP [12] or any Boolean ARM algorithm. We

use all the 27 nutrients with T10I4D100K dataset. Figure 2 shows the difference between the number of large itemsets generated from the previous method and the new FHARM approach using different fuzzy support values. The number of large itemsets increases as the minimum support decreases, naturally. In the figure, FHARM-1 uses Fuzzy HARM without normalization while FHARM-2 uses Fuzzy HARM [12] approach with normalization.

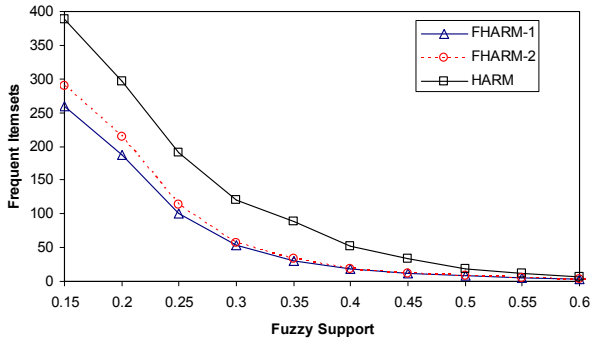

Figure 2: Number of frequent Itemsets comparison

From the results, it is clear that the approach with normalization produces more frequent itemsets (or even rules) than the converse. This is because during the normalization process, we average the fuzzy degree of fuzzy sets thus making the data more dissimilar and consequently more rules. The problem of producing many rules is easily handled by increasing the fuzzy threshold and all three algorithms become more effective as shown in figure 2.

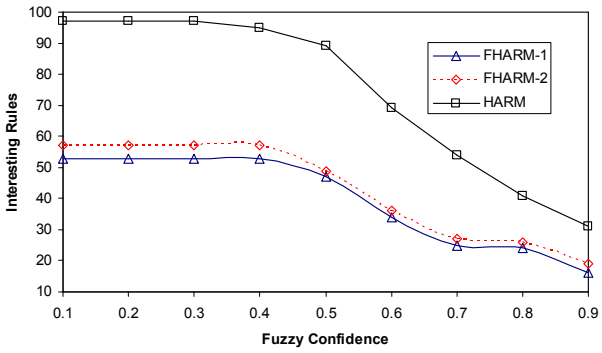

Figure 3: Number of Interesting Rules comparison

Figure 3 and Figure 4 shows the number of interesting rules using user specified fuzzy confidence and fuzzy correlation values respectively. In both cases, the number of interesting rules is less as compared to the number of rules in figure 2. Correlation has not been applied to HARM algorithm due to the boolean data and so only FHARM-1 and FHARM-2 approaches have been shown in figure 4.

Figure 4 presents more interesting rules than figure 3 because it uses the correlation value for evaluation of interestingness between the antecedent and the consequent. The experiments show that normalization before applying correlation yields significantly less rules. In addition, the novelty of the approach is in being able to analyse nutritional content of itemsets or rules. Some interesting rules produced by our approach are as follows:

IF *Protein* intake is *Ideal* THEN *Carbohydrate* intake is *low.*

IF *Protein* intake is *Low* THEN *Vitamin A* intake is *High.*

IF *Protein* intake is *High* AND *Vitamin A* intake is *Low* THEN *Fat* intake is *High.*

Depending on expert analysis, these rules are useful in analysing customer buying behaviour concerning their nutrition.

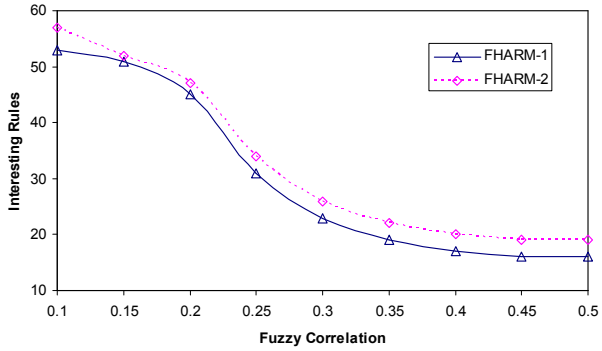

Figure 4: Number of Interesting Rules by FHARM

B. Experiment Two: (Performance Measures)

In this experiment we will show the performance measure of our new approach by varying the number of attributes and the size of data with and without normalization. The support threshold is 0.20, confidence is 0.6 and correlation value is 0.5.

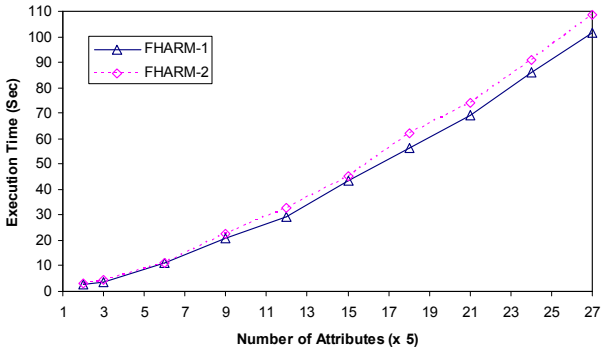

Figure 5: Performance Measures: Number of Attributes

Figure 5 shows the execution time of our algorithm with different number of attributes. Each attribute has 5 intervals or fuzzy sets, so for 3 attributes (15 columns), while 27 attributes means 135 columns. Execution time increases as we increase the number of attributes. Both algorithms have similar timings while the number of rules also increases with more attributes but fixed transactions. It is intuitive that using more attributes increases the problem's dimension.

Figure 6 below shows the execution times. We partitioned the T10I4D100K dataset into 10 equal partitions for this experiment and named them as 10K, 20K, … 100K in order to show the algorithm performance with different database sizes. For this experiment we use all 27 nutrients and set support threshold to 0.3, confidence to 0.6 and correlation value to 0.5. As data size increases, execution time and the number of rules increase. Intuitively, using normalization

results in fewer and interesting rules being generated. But similarly, the execution time increases too.

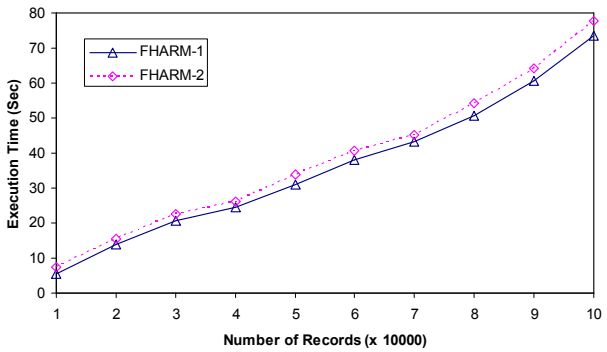

Figure 6: Performance Measures: Number of Records

From the experiments above it is shown that the fuzzy algorithms scale quite linearly to any given data sets and effective in generating fewer and interesting rules.

VII. CONCLUSION AND FUTURE WORK

In this paper, we extended the algorithm in [17] and presented a new Fuzzy HARM (FHARM) algorithm for extracting healthy buying patterns (HBP) from customer transactions. Projections are made on input data into edible attributes to find fuzzy association rules using nutrients actual membership degrees. Standard health information for each nutrient is provided as fuzzy RDA data. Fuzzy support and confidence as well as correlation are used as interestingness measures. A user can extract interesting HBP rules from the transactions or from a given rule from the rule base.

In future, we intend to evaluate our approach on real and larger customer data. Also we will compare performance of our proposed algorithm with other common fuzzy ARM algorithms [10, 15] and involve expert knowledge in evaluating the real value of HBP patterns in health terms. Overall, the approach presented here is effective and efficient for analysing HBP rules.

REFERENCES

1. Agrawal R. and Srikant R., Quest Synthetic Data Generator. IBM Almaden Research Center, http,//www.almaden.ibm.com/cs/projects/iis/hdb/Projects/data_mining/datasets/data/assoc.gen.tar.Z
2. Bodon, F., "A Fast Apriori Implementation," *Proc. IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, Vol. 90, (2003).
3. Lee, C.-H., Chen, M.-S., Lin, C.-R. , "Progressive Partition Miner, An Efficient Algorithm for Mining General Temporal Association Rules," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, (2003), 1004 - 1017.
4. Chen, G. and Wei, Q., "Fuzzy Association Rules and the Extended Mining Algorithms," *Information Sciences-Informatics and Computer Science, An International Journal archive*, Vol. 147, No. (1-4), (2002), 201 - 228.
5. Au, W-H. and Chan, K. , "Farm, A Data Mining System for Discovering Fuzzy Association Rules." *In Proc. 18th IEEE Conf. on Fuzzy Systems*, (1999), 1217-1222.
6. Srikant, R. and Agrawal, R., "Mining Quantitative Association Rules in Large Relational Tables." *In Proc. of ACM SIGMOD Conf. on Management of Data*. ACM Press, (1996), 1-12.
7. Dubois, D., Hüllermeier, E. and Prade, H., "A Systematic Approach to the Assessment of Fuzzy Association Rules," *Data Mining and Knowledge Discovery Journal*, Vol. 13, No. 2, (2006), 167 – 192.
8. Xie, D. W., "Fuzzy Association Rules discovered on Effective Reduced Database Algorithm," *In Proc. IEEE Conf.on Fuzzy Systems*, 2005.
9. He, Y., Tang, Y., Zhang, Y-Q. and Synderraman, R., "Adaptive Fuzzy Association Rule Mining for Effective Decision Support in Biomedical Applications," *International Journal Data Mining and Bioinformatics*, Vol. 1, No. 1, (2006), 3-18.
10. Gyenesei, A., "A Fuzzy Approach for Mining Quantitative Association Rules," *Acta Cybernetical*, Vol. 15, No. 2, (2001), 305-320.
11. J. Paetz, "A Note on Core Regions of Membership Functions", Proc. of the 2nd Europ. Symp. on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems, Albufeira, Portugal (2002) pp. 167-173.
12. F. Coenen, Leng, P., Goulbourne, G., "Tree Structures for Mining Association Rules," *Journal of Data Mining and Knowledge Discovery*, Vol. 15, (2004), 391-398.
13. Wang, C. and Tjortjis, C., "PRICES, An Efficient Algorithm for Mining Association Rules," *In Proc. 5th Conf. Intelligent Data Engineering Automated Learning*, Lecture Notes in Computer Science Series, Vol. 3177, Springer-Verlag, (2004), 352-358.
14. M. Delgado, N. Marin, D. Sanchez, and MA. Vila, "Fuzzy Association Rules, General Model and Applications", *IEEE Transactions on Fuzzy Systems*, 11(2),214–225, 2003.
15. Kuok, C., Fu, A., & Wong, H., "Mining Fuzzy Association Rules in Databases". *ACM SIGMOD Record*, 27, (1998), 41-46.
16. B. Xu, J. Lu, Y. Zhang, L. Xu, H. Chen and H. Yang. "Parallel Algorithm for Mining Fuzzy Association Rule", *Int'l Conf. on Cuberworld*, Singapore, 2003.
17. Muyeba M, M. Sulaiman Khan, M. Malik, Z. and Tjortjis, C. "Towards Healthy Association Rule Mining (HARM), A Fuzzy Quantitative Approach", *In Proc. 7th Conf. Intelligent Data Engineering Automated Learning*, Lecture Notes in Computer Science Series, Vol. 4224, Springer-Verlag, (2006), 1014-1022.
18. R. Agrawal, T. Imielinski, and A. Swami. "Mining Association Rules Between Sets of Items in Large Databases" *In Proc. 12th ACM SIGMOD Int'l Conf. on Management of Data*, (1993), 207-216
19. .L. Dong and C. Tjortjis, "Experiences of Using a Quantitative Approach for Mining Association Rules," *Lecture Notes in Computer Science Series*, Vol. 2690, Springer-Verlag, (2003), 693-700.