

Towards Healthy Association Rule Mining (HARM): A Fuzzy Quantitative Approach

Maybin Muyebe¹, Sulaiman M. Khan¹, Zarrar Malik², Christos Tjortjis²

¹ Liverpool Hope University, School of Computing, Liverpool, UK

² University of Manchester, School of Informatics, UK

{ muyebam@hope.ac.uk, m_sulaiman78@yahoo.com,
Zarrar.Malik@postgrad.manchester.ac.uk,
Christos.Tjortjis@manchester.ac.uk }

Abstract. Association Rule Mining (ARM) is a popular data mining technique that has been used to determine customer buying patterns. Although improving performance and efficiency of various ARM algorithms is important, determining Healthy Buying Patterns (HBP) from customer transactions and association rules is also important. This paper proposes a framework for mining fuzzy attributes to generate HBP and a method for analysing healthy buying patterns using ARM. Edible attributes are filtered from transactional input data by projections and are then converted to Required Daily Allowance (RDA) numeric values. Depending on a user query, primitive or hierarchical analysis of nutritional information is performed either from normal generated association rules or from a converted transactional database. Query and attribute representation can assume hierarchical or fuzzy values respectively. Our approach uses a general architecture for Healthy Association Rule Mining (HARM) and prototype support tool that implements the architecture. The paper concludes with experimental results and discussion on evaluating the proposed framework.

Keywords: Association rules, healthy patterns, fuzzy rules, primitive and hierarchical queries, nutrients.

1 Introduction

Association rules (ARs) [1] have been widely used to determine customer buying patterns from market basket data. Most algorithms in the literature have concentrated on improving performance through efficient implementations of the modified *Apriori* algorithm [2], [3]. Although this is an important aspect in large databases, extracting health related information from association rules or databases has mostly been overlooked. People have recently become “healthy eating” conscious, but largely they are unaware of qualities, limitations and above all, constituents of food. For example, how often do people who buy baked beans bother with nutritional information other than looking at expiry dates, price and brand name? Unless the customer is diet conscious, there is no explicit way to determine nutritional requirements and

consumption patterns. As modern society is concerned with health issues, association rules can be used to determine healthy buying patterns by analysing product nutritional information, here termed Healthy Association Rule Mining (HARM), using market basket data. The term Healthy Buying Patterns (HBP) is introduced and signifies the level of nutritional content in an association rule per item.

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

2 Background and Related Work

In almost all AR algorithms, thresholds (both confidence and support) are crisp values. This support specification may not suffice for our approach and we need to handle linguistic terms such as “low protein” etc. in queries and rule representations.

Fuzzy approaches [4], [5] deal with quantitative attributes [6] by mapping numeric values to boolean values. A more recent overview is given in [7]. Little attention has been given to investigating healthy buying patterns (HBP) by analysing nutrition consumption patterns. However [8] presents fuzzy associations by decreasing the complexity of mining such rules using a reduced table. The authors also introduce the notion of mining for nutrients in the antecedent part of the rule but it is not clear how the fuzzy nutrient values are dealt with and consequently how membership functions are used. Nutrient analysis is therefore more complex a process than mere search for element presence. Our approach determines whether customers are buying healthy food, which can easily be evaluated using recommended daily allowance (RDA) standard tables. Other related work dealing with building a classifier using fuzzy ARs in biomedical applications is reported in [9].

3 Problem Definition

The problem of mining fuzzy association rules is given following a similar formulation in [10]. One disadvantage discussed, is that discretising quantitative attributes using interval partitions brings 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.

Given a database D of transactions and items $I = \{i_1, i_2, \dots, i_m\}$, 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.

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 , $0 \leq m_{i_j}(l, v) \leq 1$. For each transaction $t \in E$, a normalization process to find significance of an item's contribution to the degree of support of a transaction is given by equation (1):

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

The normalisation process ensures fuzzy membership values for each nutrient are consistent and are not affected by boundary values. To generate fuzzy support (FS) value of an item set X with fuzzy set A , we use equation (2):

$$FS(X, A) = \frac{\sum_{t_i \in T} \pi_{x_j \in X} m_{x_j}(a_j \in A.t_i.x_j)}{|E|} \quad (2)$$

A quantitative rule represents each item as <item, value> pair. For a rule $\langle X, A \rangle \rightarrow \langle Y, B \rangle$, the fuzzy confidence value (FC) where $X \cup Y = Z, A \cup B = C$ is given by equation (3):

$$FC(\langle X, A \rangle \rightarrow \langle Y, B \rangle) = \frac{\sum_{t_i \in T} \pi_{z_c \in X} m_{z_j}(c_j \in C.t_i.z_j)}{\sum_{t_i \in T} m_{x_j}(a_j \in A.t_i.x_j)} \quad (3)$$

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

$\bigcup_{k=1}^p i_j^k, 1 \leq j \leq m$. 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 $\sum \frac{|i_j^5|}{3}, j \in \{3,4,7\}$. These values are then aggregated

into an RDA table with a schema of nutrients (see table 2, in 4.1) 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 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}.

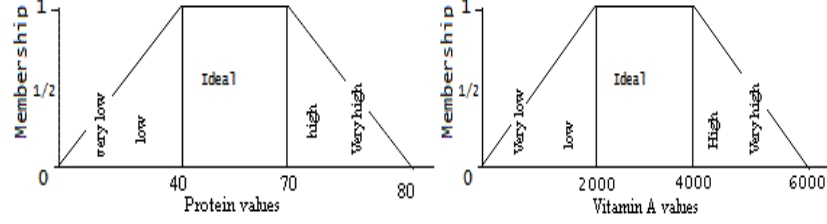


Fig. 1. Fuzzy membership functions

$$\mu(x, \alpha, \beta, \gamma, \delta, \theta) = \begin{cases} 0, & \delta < x < \alpha \\ \frac{(\alpha - x)\theta}{(\alpha - \beta)}, & \alpha \leq x \leq \beta \\ \theta, & \theta = 1 \\ \frac{(\delta - x)\theta}{(\delta - \gamma)}, & \gamma \leq x \leq \delta \end{cases} \quad (4)$$

Examples of fuzzy membership functions for some nutrients are shown in figure 1 (Protein and Vitamin A). 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. Equation 4 [11] represents all nutrient membership functions with input range of ideal values and the initial and final range of all values.

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.

4 Proposed Methodology

The proposed methodology consists of various HARM queries, each of which is evaluated using fuzzy sets for quantitative attributes as mentioned earlier. We can use any Apriori-type algorithm to generate rules but in this case Apriori TFP (Total From Partial) ARM algorithm is used as it is efficient and readily available to us. Apriori TFP stores large items in a tree and pre-processes input data to a partial P tree thus making it more efficient than Apriori and can also handle data of duplicate records. We have discovered three techniques to obtain HBPs as described in the next sections.

4.1 Normal ARM Mining

To mine from the transactional file (table 1), input data is projected into edible database on-the-fly thereby reducing the number of items in the transactions and possibly transactions too. The latter occurs because some transactions may contain non-edible items which are not needed for nutrition evaluation. This new input data is converted into an RDA transaction table (table 2) 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 {1, 2, 3, 4, 5} for the first item or nutrient, {6, 7, 8, 9, 10} for the second nutrient and so on. The first nutrient, protein (Pr), is coded 1 to 5 and based on equation 4, we can determine the value 20 as “Very Low” or VL etc. Thus nutrient Pr has value 1 in table 3. The encoded data (table 3) 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.

Table 1. Transaction file

TID	Items
1	X, Z
2	Z
3	X, Y, Z
4	..

Table 2. RDA transactions

TID	Pr	Fe	Ca	Cu
1	20	10	30	60
2	57	70	0	2
3	99	2	67	80
4

Table 3. Fuzzy transactions

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

Table 4. 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

The other approach is to convert RDA transactions (table 2) to linguistic values 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 (table 4), will have repeated fuzzy values {very low, low, ideal, high, very high} for each nutrient present in every item of that transaction. Table 4 actually shows only two nutrients. A data structure is then used to store these values (linguistic value and degree of membership) and large itemsets are 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 obtain ARs in the normal way with HBP values.

4.2 Rule Query on Nutrient Associations

To mine a specific rule, $X \rightarrow Y$, for nutritional content, the rule base (table 5) is scanned first for this rule and if found, converted into an RDA table (table 6)

otherwise, the transactional database is mined for this specific rule. The latter involves projecting the database with attributes in the query, thus reducing the number of attributes in the transactions, and mining as described in 4.1.

In the former case, HBP is calculated and the rule stored in the new rule base with appropriate support, for example [proteins, ideal] \rightarrow [carbohydrates, low], 35%. A rule of the form “Diet Coke \rightarrow Horlicks, 24%” could be evaluated to many rules including for example, [Proteins, ideal] \rightarrow [Carbohydrates, low], 45%; where, according to rule representations shown in section 3, X is “Proteins”, A is “ideal” and Y is “Carbohydrates”, B is “low” etc. The same transformation to an RDA table occurs and the average value per nutrient is calculated before conversion to membership degrees or linguistic values. Using equations 1, 2, 3 and 4, we evaluate final rules with HBP values expressed as linguistic values. The following example shows a typical query as described in 4.1 where TID is transaction ID, X,Y, Z are items and P (protein), Fe (Iron), Ca (calcium), Cu (Copper) are nutritional elements and support of N% is given:

Table 5. Rule base

Rules	Support
X \rightarrow Y	24%
Y \rightarrow Z	47%
X,Y \rightarrow Z	33%
..	..



Table 6. RDA table and HBP rule

	Pr	Fe	Ca	Cu	..
X \rightarrow Y	20	10	30	60	..

X \rightarrow Y [Proteins, Very Low] \rightarrow [Carbohydrates, Low], s=45%, c=20%;

4.3 Hierarchical Rule Query

To make the system usable by a variety of users, hierarchical queries may be needed and a tree parsing algorithm can be used to obtain leaf nodes or concepts of the hierarchical query terms can be retrieved. After obtaining leaf terms, the mining algorithm proceeds as in section 4.2. For example, a hierarchical rule query such as:

$$\text{Vegetable (V)} \rightarrow \text{Meat (P)}$$

where Vegetable is parsed to lettuce, cabbage etc. and meat to beef, liver etc. can be a typical query for other types of users.

5 Architecture

The proposed framework has a number of components in the architecture (see figure 2). Firstly, a user query is given for a specific task and the HARM Manager through the Query Detector determines the query type. If it is a query type described in section 4.1, the edible filter is activated and an RDA table generated for the given transactions for edible items. The Data Mining (DM) module then invokes an

appropriate association rule algorithm. If the query is as in 4.2, then the RDA Converter is activated to generate RDA transactions for that rule and then an algorithm in the DM module is used. In our approach, we have thought it useful, for future use, to keep generated RDA transactions so that we can test other AR algorithms.

The fuzzy module involves tree parsing of hierarchical query terms (items) and determining fuzzy sets for these items' leaf concepts which become predefined rules. After filtration and RDA conversion is done, the query is then passed to the DM module where an appropriate algorithm is run for a particular query. Rules are generated and stored in the rule base.

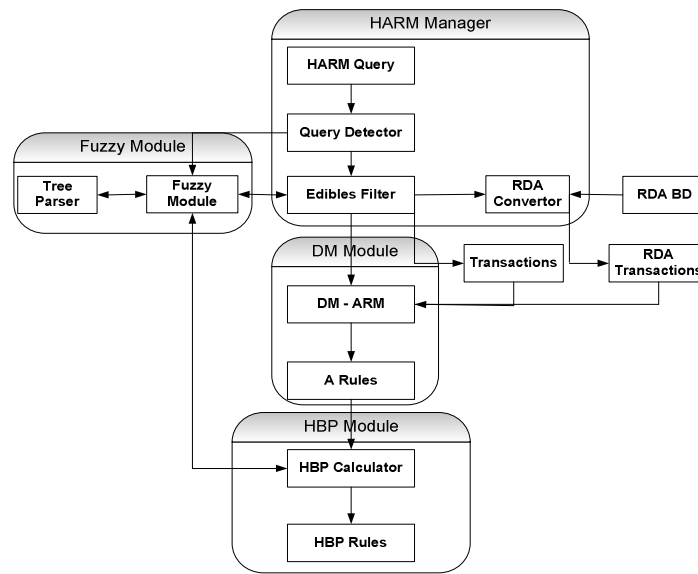


Fig. 2. HARM Architecture

After generating the rule base or finding the support and confidence for predefined rules, the rules are passed to the HBP Module where the HBP calculator is activated which uses fuzzy functions to evaluate the HBP strength of given rules as outlined in section 4.2.

6 Experimental Results

In order to show the defined frameworks effectiveness, we performed experiments using the prototype system with synthetic data (1 million transactions with 30 edible items out of 50 items) and used a real nutritional standard RDA table to derive fuzzy values. Our choice of association rule algorithm [12] was based on efficiency and availability. We also implemented the algorithms for analysing rule queries and

calculating fuzzy support and fuzzy confidence. For missing nutrient values or so called “trace” elements, the fuzzy function evaluated zero degree membership. We run AprioriTFP on the data to produce a rule base. Some of the rule queries are as follows:

Rule 1: Milk → Honey, Support=29%

The rule is evaluated accordingly (see 4.2) as

HBP is

44% - Very Low in [Calcium Cholesterol Fats Iodine Magnesium Manganese Phosphorus Sodium VitaminA VitaminC VitaminD VitaminK]
3% - Low in [VitaminB12]
14% - Ideal in [Fiber Protein VitaminB6 Zinc]
7% - High in [Niacin VitaminE]
29% - Very High in [Biotin Carbohydrate Copper Folacin Iron Riboflavin Selenium Thiamin]

Rule 2: Cheese, Eggs → Honey, Support=19%

HBP is

37% - Very Low in [Calcium Fats Iodine Magnesium Phosphorus VitaminA VitaminB12 VitaminC VitaminD VitaminK]
3% - Low in [Carbohydrate]
22% - Ideal in [Manganese Protein Sodium VitaminB6 VitaminE Zinc]
3% - High in [Cholesterol]
33% - Very High in [Biotin Copper Fiber Folacin Iron Niacin Riboflavin Selenium Thiamin]

Rule 3: Jam → Milk, Support=31%

HBP is

48% - Very Low in [Calcium Cholesterol Fats Iodine Iron Magnesium Phosphorus] etc.

It is surprising to see that for most rules (at least these shown here), calcium purchases from calcium rich products like milk and cheese are very low. Contrary, Biotin (Vitamin H, rules 1 and 2) deficiency that causes cholesterol, loss of appetite, hair loss etc is very high possibly because it is found in egg yolks and milk (dry skimmed). These inferences could be useful in real data applications.

7 Conclusion and future work

In this paper, we presented a novel framework for extracting healthy buying patterns (HBP) from customer transactions by projecting the original database into edible attributes and then using fuzzy association rule techniques to find fuzzy rules. In this new approach, a user can formulate different types of queries to mine ARs either from the transactions, or from a given rule from the rule base or using a

hierarchical query. Standard health information for each nutrient is provided as fuzzy data to guide the generation and evaluation of the rules.