A Framework for Mining Fuzzy Association Rules from Composite Items

Muhammad Sulaiman Khan[1]

Dr Maybin Muyeba[2]

Dr Frans Coenen[3]

[1] Liverpool Hope University

[2] Manchester Metropolitan University

[3] The University of Liverpool

IFIP 2008, Milano, Italy

Outline of the Presentation

Organised as follows:

- Introduction
  - Classical Association Rule Mining (ARM)
  - Quantitative Association Rule Mining
  - Fuzzy Association Rule Mining (FARM)
- Problem definition
- Methodology
- Example
- Conclusion & Further work

Introduction

- **Association Rule Mining (ARM)**
  - Data Mining Technique for finding “interesting” patterns in binary valued data sets.
  - Patterns usually translated into Association Rules (ARs) of the form

    $$X \rightarrow Y$$

    where X and Y are item sets.
  - ARM algorithms usually operate using the support-confidence frame work, and utilise the Downward Closure Property (DCP) of itemsets.

Quantitative Association Rule Mining

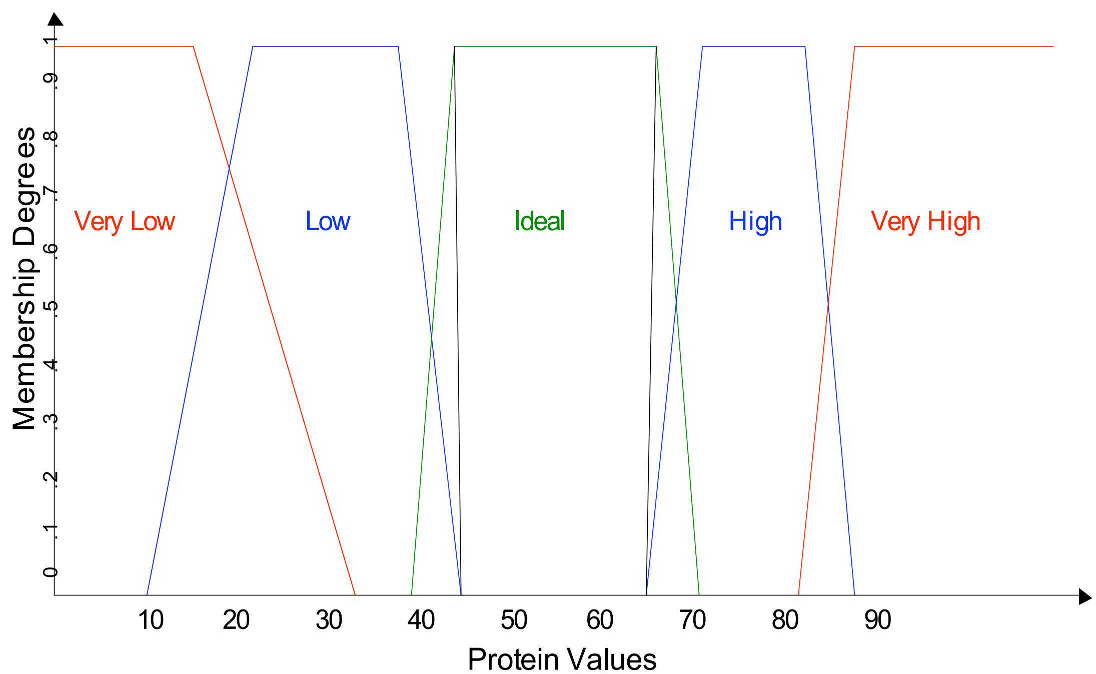
- **Quantitative ARM**
  - Applied to non-boolean data.
  - Data is discretized.
  - In the case of numeric quantiutative data items this causes what is known as the “crisp boundary” problem.

Fuzzy Association Rule Mining (FARM)

- Fuzzy sets used to resolve the Crisp Boundary problem by providing a smooth change between boundaries.
- Fuzziness is defined by a membership mapping function.

$$\mu(x) : A \rightarrow [0,1], x \in A$$

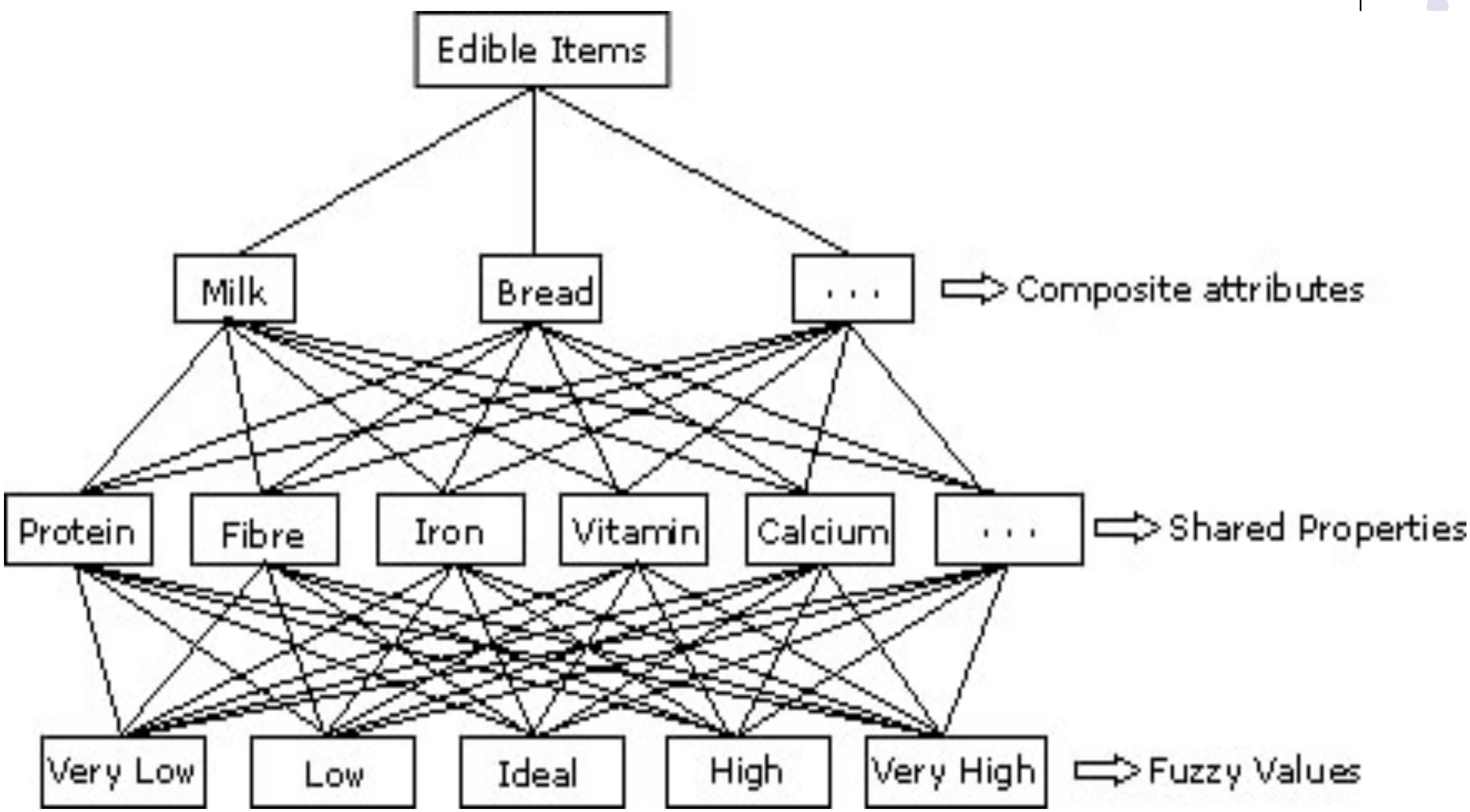
- Example (Trapezoidal membership function):



FARM using composite attributes

- FARM extended to composite Attributes
- Composite Attributes
  - Objects (items) with collections of properties (set of values).
  - Properties can be quantitative or categorical.
  - Properties are shared across the attribute set.
  - Quantitative properties can be fuzzified into several ranges (fuzzy sets).

FARM using composite attributes (Example)



Problem definition

- Given a Dataset *D* consisting of a set of transaction $t=\{t_1,t_2,t_3,..,t_n\}$, a set of composite items $I=\{i_1,i_2,i_3,..,i_{|I|}\}$ *and a set of properties* $P=\{p_1,p_2,p_3,..,p_m\}$.
- Each transaction t_i is subset of *I*, and each item $t_i[i_j]$ is a subset of *P*.
- Thus each item i_j will have associated with it a set of numeric values corresponding to the set *P*, i.e. $t_i[i_j]=\{v_1,v_2,v_3,..,v_m\}$.

Problem Definition

- Example

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

$D=\{t_1, t_2, t_3, t_4\}$

$I=\{a, b, c, d\}$

$P=\{1,2,3,4,5,6\}$

Problem Definition

- Property Dataset
  - D is initially transformed into a Property dataset D^P.
  - D^P consists of "Property Transactions" $T^P=\{t^P{}_1,t^P{}_2,t^P{}_3,..,t^P{}_n\}$.
  - Each transaction $t^P{}_{i\,,}$ is subset of $P=\{p_1,p_2,p_3,..,p_m\}$.
  - The value for each Property attribute $t^P{}_i[P_j]$ is obtained by summing the numeric values for all p_j in t_i. *Thus*

$$t^p{}_i[p_j] = \frac{\sum_{j=1}^{|t_i|} t^i[i_j[v_k]]}{|t_i|}$$

Problem Definition

- Fuzzy Dataset
  - D^P is further transformed into a fuzzy dataset $D^/$.
  - A fuzzy dataset $D^/$ consists of fuzzy transactions $T^/=\{t^/_1,t^/_2,t^/_3,..,t^/_n\}$ *and fuzzy property attributes* $P^/$.
  - Each $P^/$ has a number of fuzzy sets associated with it, identified by a set of linguistic labels L=$\{l_1,l_2,l_3,..,l_{|L|}\}$ e.g. {small, medium, large}.
  - Each property attribute $t^P_i[P_j]$ is associated (to some degree) with several fuzzy sets, with a membership degree in the range [0,1].
  - Membership degree indicates the correspondence between the value of a given $t^p_i[p_j]$ and the set of *fuzzy linguistic labels*.

Problem Definition

- Composite Item Value Table
  - A composite item value table is a “look-up” table that allows us to get property values for specific items.

- Properties Table
  - A properties table is a table that maps all possible values for each property attribute $t^{P}_{i}[P_{j}]$ onto fuzzy/overlapped ranges.

Problem definition

- Fuzzy Normalisation Process (total membership degree value for properties to add up to 1)
  - The process of finding the contribution to the fuzzy support value, $m^{/}$, for individual property attributes $t^{/}_{i}[p_j[l_k]]$ such that a partition of unity is guaranteed.

$$t'_i[p_j[l_k]] = \frac{\mu(t^p{}_i[p_j[l_k]])}{\sum_{x=1}^{|L|} \mu(t^p{}_i[p_j[l_x]])}$$

TID	VL	L	ID	H	VH	…
1	0.0	0.0	0.0	1.0	.32	…
2	.83	.38	0.0	0.0	0.0	…
3	…	…	…	…	…	…

TID	VL	L	ID	H	VH	…
1	0.0	0.0	0.0	.76	.24	…
2	.69	.31	0.0	0.0	0.0	…
3	…	…	…	…	…	…

Problem definition

- Fuzzy Support
  - Fuzzy support is calculated as

$$FS(A) = \frac{Sum\ of\ votes\ satisfying\ A}{Number\ of\ records\ in\ T'}$$

$$\text{votes for } t_i \text{ satisfying } A = \sum_{i=1}^{i=n} \prod_{\forall [i[l]] \in A} t'[i_j[l_k]]$$

$$FS(A) = \frac{\sum_{i=1}^{i=n} \prod_{\forall [i[l]] \in A} t'[i_j[l_k]]}{n}$$

Problem definition

- Fuzzy Confidence
  - Fuzzy confidence (FC) is calculated in the same manner that confidence is calculated in traditional ARM.
  - Fuzzy confidence is calculated as:

$$FC(A \to B) = \frac{FS(A \cup B)}{FS(A)}$$

Methodology

- Data Transformation
  - Transformation of raw dataset *T* into property dataset T^p.
  - Transformation of property dataset T^p into a database containing fuzzy extensions $T^{/}$.
  - Normalization of fuzzy dataset.
  - Candidate Generation i.e. search for all fuzzy frequent itemsets that have support higher than user specified threshold.
  - Use frequent itemsets to generate all possible rules using fuzzy confidence or fuzzy correlation interestingness measures.

Example Application

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	…
VitaminA	0	15	150	200	150	200	300	400	300	350	440	500	440	490	550	600	550	600	…
Zinc	0	.8	8	10	8	10	15	20	15	20	30	40	30	40	46	50	46	50	…
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

(a) Raw data (T)

TID	Items
1	X, Z
2	Z
3	X,Y, Z
4	…

(b) Property data set (T^P)

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

Example Application

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	…	…	…	…	…	…	…	…	…	…	…

Experimental Results

- Some example fuzzy rules produced by our approach (30% support, 50% confidence and 25% correlation) 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.*
- It is suggested that these rules are useful in analysing customer buying behavior concerning their nutrition.

Conclusion & Further Work

- We have presented an approach for extracting hidden information from composite items.
- We showed that with such items, common properties can be defined as quantitative itemsets themselves, which are transformed into fuzzy sets.
- Overall, the approach presented is effective and efficient for analysing databases with composite items.
- Further work will evaluate our approach on real and larger datasets and compare real performance with other common fuzzy ARM algorithms.
- There is potential to apply this to other applications with composite items or attributes even with varying fuzzy sets between attributes e.g. image analysis and inventory control database.
- We are expanding our work with the possibilities to extend it for Fuzzy Utility and Weighted Association Rule Mining.