A Framework for Mining Fuzzy Association Rules from Composite Items

Muhammad Sulaiman Khan [1]
Dr Maybin Muyeba [1]
Dr Frans Coenen [2]

[1]Liverpool Hope University
[2]Liverpool University

Outline of the Presentation

Organised as follows:

- Introduction
  - Classical Association Rule Mining (ARM)
  - Quantitative Association Rule Mining
  - Fuzzy Association Rule Mining (FARM)
- Background & Related Work
- Problem definition
- Methodology & Application
- CFARM Algorithm
- Experimental Results
- Conclusion & Further work

Introduction

- **Association Rule Mining**
  - Data Mining Technique
  - Determine customer buying Patterns from market basket data/Transactions.
  - Association rules are of the form

    X → Y

  where X and Y are item sets and $X \cap Y = f$

Association Rule Mining

- **Support**

  Supp (X → Y) = Supp (X U Y)

- **Confidence**

  Conf (X → Y) = Supp (X U Y) / Supp (X), a conditional probability

- **Downward Closure Property (DCP)**
  - Subsets of a frequent set are also frequent, e.g. if {A,B,C} is a frequent set then {A,B}, {A,C} and {B,C} will also be frequent.

Quantitative Association Rule Mining

- **Quantitative ARM**
  - Applies on non-boolean and relational databases
  - Determine rules of the form:

    if (X is A) then (Y is B)

  where X and Y are attributes in a database and A and B are the discretised values of these attributes.
  For example:

    if <Age is Young> then <Salary is Low>

Quantitative Association Rule Mining (Fuzziness)

RecordID	Age	Income
100	31	26000
200	16	10000
300	45	60000
400	_	_

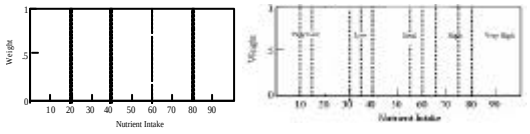
Age	Intervals
0-25	Young
26-40	Middle
41-60	Old
60-above	Very Old

Income	Intervals
0-12K	Low
13K-30K	Medium
31K-45K	High
46K-above	Very High

TID	Young	Middle	old	V Old	Low	Medium	High	V High
100	0	1	0	0	0	1	0	0
200	1	0	0	0	1	0	0	0
300	0	0	1	0	0	0	0	1
400	_	_	_	_	_	_	_	_

Fuzzy Association Rule Mining (FARM)

- Issues from partitioning quantitative attributes
  - Interval Partitions give boundary Problems to quantitative attri butes using usual support measure (Supp)
  - To resolve, borders can overlap but this might over emphasise some intervals (Fuzziness is evident).
  - Example: Nutrition content of food – a protein or vitamin etc.

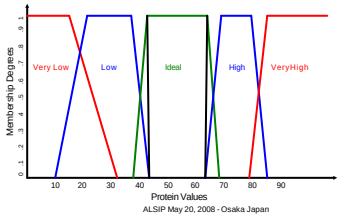


Fuzzy Association Rule Mining (FARM)

- Fuzzy sets used to resolve this by Providing a smooth change between boundaries using normalisation process (see later!).
- Firstly, fuzziness is defined by a membership mapping function

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

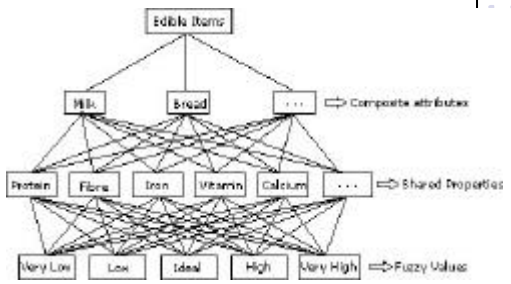
- Trapezoidal membership function can be derived (example above)



FARM using composite attributes

- FARM extended to composite Attributes
- Composite Attributes
  - Objects with different Properties
  - Properties can be quantitative and categorical
  - Attributes share the same Properties with other attributes
  - Quantitative Properties can be discretised into several ranges (fuzzy sets)
  - ARM usually on all data sets (edible, non-edible)

FARM using composite attributes



Background& Related Work

- From literature the only works that mention "Composite ARM" was Presented in 1997 & 2006.
  - Ye, X.; Keane, J.A. "Mining association rules with composite items", IEEE International Conference on Systems, Man, and Cybernetics, PP 1367 – 1372, Oct 1997.
  - Ke Wang, James N. K. Liu, Wei-min Ma, "Mining the Most Reliable Association Rules with Composite Items," *icdmw*, PP. 749-754, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 2006.
- In these works, Composite items expressed as combination of several attributes
  - Composite means put B and C to make a new item {BC}
  - If itemset {A,B} and {B,C} are not frequent (large)
    - B→A and C→A will not be generated
    - Rule {BC}→{A} may be generated.

Problem definition

- Given a Dataset *D* consist of 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. $t_i[i_j]=\{v_1,v_2,v_3,..,v_m\}$.
- *The "k^{th}" property value for the "j^{th}" item in "ith" transaction is given by* $t_i[i_j[v_k]]$.

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}>}

$D=\{t_1, t_2, t_3, t_4\}$
$I=\{a, b, c, d\}$
$P=\{x, y, z\}$

Problem Definition

- Property Dataset
  - D is initially transformed into a Property dataset D^P.
  - D^P consists of Property transactions $T^P=\{t^P_1, t^P_2, t^P_3, .., t^P_n\}$.
  - For 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 aggregating 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 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, e.g. [0-10], [7-19], [15-30]

Problem definition

- Fuzzy Normalisation Process
  - 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.

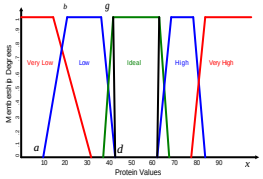
$$t^{/}_i[p_j[l_k]] = \frac{m(t^p{}_i[p_j[l_k]])}{\sum_{x=1}^{|L|} m(t^p{}_i[p_j[l_x]])}$$

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

- Membership Function (Trapezoidal)
  - Membership degree to a particular fuzzy set (described by linguistic label), $t^{/}_i[p_j[l_k]]$ is determined by a membership function $m(x): A \rightarrow [0,1], x \in A$



$$m(x,a,b,g,d) = \begin{cases} 0, d < x < a \\ \dfrac{(x-a)}{(b-a)}, a \le x \le b \\ \dfrac{(d-x)}{(d-g)}, g \le x \le d \\ 1, b < x < g \end{cases}$$

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:

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

Problem definition

- **Fuzzy Correlation Measure**
  - Fuzzy Confidence does not take into account *FS(B)*.
  - The Fuzzy Correlation (*FCORR*) addresses this.
  - Similar to statistical correlation but different in meaning.

$$FCorr(A \rightarrow B) = \frac{Cov(A, B)}{\sqrt{Var(A) \times Var(B)}}$$

$$Cov(A, B) = E[C] - E[A] \times E[B] \quad \text{where } C = A \cup B$$

$$Cov(A, B) = FS(C) - FS(A) \times FS(B)$$

Problem Definition

The variance of A and B can be obtained as follows:

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

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

where

$$E[X]^2 = FS(X)^2 = \left(\frac{\sum_{i=1}^{i=n} \prod_{\forall [i[l]] \in X} t'[i_j[l_k]]}{n} \right)^2$$

$$E[X^2] = FS(X^2) = \frac{\sum_{i=1}^{i=n} \prod_{\forall [i[l]] \in X} (t'[i_j[l_k]])^2}{n}$$

Proposed 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.

Proposed Methodology

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

TID	Pr	Fe	Ca	..
1	53	41	77	..
2	27	19	30	..
3	99	2	67	..
4	..	..	..	..

TID	VL	L	Ideal	H	VH	VL	L	Ideal	H	VH	..
1	0.0	0.0	0.9	0.1	0.0	0.0	0.4	0.6	0.2	0.8	..
2	0.4	0.6	0.0	0.0	0.0	0.5	0.5	0.0	0.0	0.0	..
3	0.0	0.0	0.0	0.7	0.3	0.0	0.0	0.3	0.7	0.0	..
4	..	..	..	..	..	..	..	..	..	..	..

CFARM Algorithm

Properties Converter(T, RDA, $I_{edibles}$)

1. $\forall$ *T in D*
2. $\forall$ *I in t*
3. $T^P = T^P \cup P_j$
4. T^P = *average Contributi on(Tc)*
5. D^P = *wrirte*(T^P)
6. *end;*

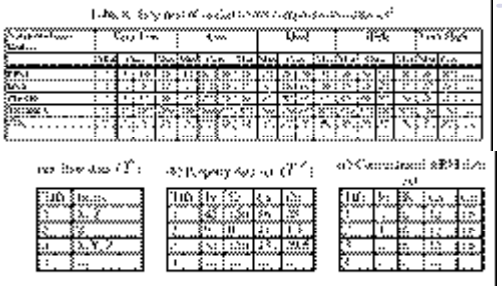
FuzzyConverter(T^P, Properties Table)

1. $\forall T^P$ *in* D^P
2. $\forall P_j$ *in* T^P
3. *fuzzyattr= getFuzzyAt tr*(P_j, *propTable*)
4. $T' = T' \cup$ (*fuzzyattr*)
5. $D' =$ *wrirte*(T')
6. *end;*

CFARM(*minsupp, minconf, mincorr, T'*)

1. $k = 0;\ C_k = f = f_k = f$
2. *do*
3. $k = k+1$
4. *if*($k == 1$)
5. $Ck =$ *Gener ateFirstCa ndidates* (T')
6. *else*
7. $Ck =$ *Gener atecandidates*(F_{k-1})
8. $\forall$ *Ck*
9. *count = Co untSupport*(C_k)
10. C_k=*PruneCandidates*(C_k,*count, minsupp*)
11. C_k = *CalcSig nificance*(C_k , *minconf*)
12. F_k =*GenerateF requentIte msets*(C_k ,*minconf*)
13. $F = F \cup F_k$
14. *while*(C_k .*count*> *k*)
15. *FuzzyCorr = CalcCert ainty*(*F , mincorr*)
16. *Output (Rules*(*F , m incorr, Fu zzyCorr*))

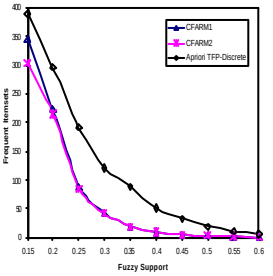
Example Application



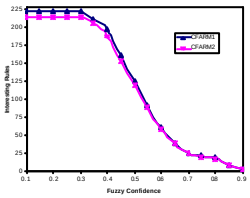
Example Application

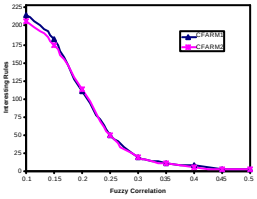
Experimental Results

- Quality Measures
  - We use all standard 27 nutrients present in edible items.
  - We use T10I4D100K dataset with 100K transactions.
  - Fuzzy Support 30%
  - Fuzzy Confidence 50%
  - Fuzzy Correlation 25%



Experimental Results

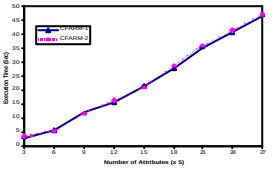


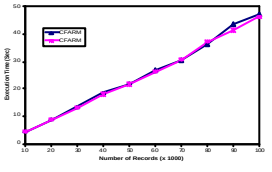


Experimental Results

- Performance Measures

  We set a support threshold to 0.30, confidence 0.5 and correlation value to 0.25.





Experimental Results

- Some interesting fuzzy rules produced by our approach with 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*.
- Depending on expert analysis from a health practitioner, these rules are useful in analysing customer buying behavior concerning their nutrition.

Conclusion & Further Work

- We have presented a novel 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.
- We use fuzzy correlation measure as interestingness measure and showed more interesting rules.
- Overall, the approach presented here 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, inventory control database.
- We are expanding our work with the possibilities to extend it for Fuzzy Utility Association Rule Mining.