A Weighted Utility Framework for Mining Association Rules

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

[1]Liverpool Hope University
[2]The University of Liverpool

Overview

Organised as follows:

- Introduction
  - Association Rule Mining (ARM)
  - Downward Closure Property (DCP)
  - Weighted ARM
- Our Contribution
  - Weighted Utility Hybrid Framework
- Methodology
- Simulated Example
- Evaluation
  - Dataset
  - Quality Measures
  - Performance Measures
- Conclusion

Liverpool Hope University

Introduction

- Data Mining
- Association Rule Mining (ARM)
- Qualitative vs Quantitative
  - Database count
  - Items' significance
  - Items' frequencies
- Standard ARM only deals with database count
- Standard AR's may contribute only a small portion of the overall company profit
- Anti-monotonic property does not hold

Introduction

Table 1. Weighted items table

ID	Item	Profit	Weight	...
1	Shirt	£10	0.1	...
2	Jean	£25	0.3	...
3	Jacket	£50	0.6	...
4	Suit	£80	0.9	...

Table 2. Customers transactions

Tid	Shirt	Jean	Jacket	Suit
1	1	1	0	1
2	0	2	1	0
3	1	1	2	1
4	1	0	1	1

[jeans → suit, 50%]

[shirt → suit, 75%]

Association Rule Mining

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

$$X \rightarrow Y$$

  - where X and Y are item sets and
  - Measures
    - **Support**: Supp $(X \rightarrow Y)$ = Supp $(X \cup Y)$
    - **Confidence**: Conf $(X \rightarrow Y)$ = Supp $(X \cup Y)$/Supp (X)

Liverpool Hope University

Downward Closure (DCP)

- 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.

  - Applications
    - Help algorithms to generate large itemsets of increasing size by adding items to itemsets that are already large.
    - we assume that if AB and BC are not frequent, then ABC and BCD cannot be frequent so we don't consider generating the supersets that contain non-frequent itemsets.

Weighted Association Rule Mining

- Standard ARM model assumes that all items have the same significance without taking account of their weight within a transaction or record.

For example rules:

A: [computer → monitor, 5%, 80%], **B:** [printer → scanner, 13%, 80%]

In standard ARM rule ***B*** is more important than rule ***A*** because rule ***B*** has higher support than rule ***A***.

But in weighted ARM with weighted settings rule ***A*** may be more important than rule ***B***, even though the former holds a lower support.

This is because those items in the first rule usually come with more profit per unit sale, but the standard ARM simply ignores this difference.

Our Contribution

- Weighted Utility ARM (Hybrid Framework)
- WUARM as extension of weighted and Utility ARM
  - Significance of itemsets
  - Frequency of itemsets
- Weighted Utility of an itemset
  - Transactional Utility:
    - It is the frequency of occurrences or quantity of an item in a transaction.
  - Item significance:
    - It is the value representing significance of an item (value, profit etc) and it holds across the dataset.
- Item sets holds DCP
- WUARM: modified Apriori algorithm

Proposed Methodology

- Item Weight $w(i_j)$
- Weighted Table $WT(I, W)$
- Item Utility $t_q(i_j, u)$
- Item Weighted Utility $t_i[(w(i_j), u)]$
- Transaction Weighted Utility $$twu(t_i) = \frac{\sum_{j=1}^{|t_i|} t_i[(w(i_j), u)]}{|t_i|}$$
- Weighted Utility Support $$wus(XY) = \frac{\sum_{i=1}^{|S|} twu(t_i)}{\sum_{i=1}^{|T|} twu(t_i)}$$

$$S = \{S \mid S \subseteq T, X \cup Y \in S\}$$

Simulation

Table 3. Weighted items table

Items *i*	Profit	Weights *w*
A	£60	0.6
B	£10	0.1
C	£30	0.3
D	£90	0.9
E	£20	0.2

Table 4. Transaction database with transactional weighted utilities of items

Items	A	B	C	D	E	*twu*
1	1	1	4	1	0	0.700
2	0	1	0	3	0	1.400
3	2	0	0	1	0	1.050
4	0	0	1	0	0	0.300
5	1	2	0	1	3	0.575
6	1	1	1	1	1	0.420
7	0	2	3	0	1	0.433
8	0	0	0	1	2	0.650
9	7	0	1	1	0	1.800
10	0	1	1	1	1	0.375
			Weighted Utility count			7.703

Table 5. Weighted utility mining comparison

#	Standard ARM	Weighted ARM	Weighted Utility ARM
1.	A (50%)	A (30%)	A (0.59)
2.	A→B (30%)	A→B (21%)	A→B (0.22)
3.	A→B→C (20%)	A→B→C (20%)	A→B→C (0.14)
4.	A→B→C→D (20%)	A→B→C→D (38%)	A→B→C→D (0.14)
5.	A→B→C→D→E(10%)	A→B→C→D→E(21%)	A→B→C→D→E (0.05)
6.	A→B→C→E (10%)	A→B→C→E (12%)	A→B→C→E (0.05)
7.	A→B→D (30%)	A→B→D (48%)	A→B→D (0.22)
8.	A→B→D→E (20%)	A→B→D→E (36%)	A→B→D→E (0.13)
9.	A→B→E (20%)	A→B→E (18%)	A→B→E (0.13)
10.	A→C (30%)	A→C (27%)	A→C (0.38)
11.	A→C→D (30%)	A→C→D (54%)	A→C→D (0.38)
12.	A→C→D→E (10%)	A→C→D→E (20%)	A→C→D→E (0.05)
13.	A→C→E (10%)	A→C→E (11%)	A→C→E (0.05)
14.	A→D (50%)	A→D (75%)	A→D (0.590)
15.	A→D→E (20%)	A→D→E (34%)	A→D→E (0.13)
16.	A→E (20%)	A→E (16%)	A→E (0.13)
17.	B (60%)	B (6%)	B (0.51)
18.	B→C (40%)	B→C (16%)	B→C (0.25)
19.	B→C→D (30%)	B→C→D (39%)	B→C→D (0.19)
20.	B→C→D→E (20%)	B→C→D→E (30%)	B→C→D→E (0.10)
21.	B→C→E (30%)	B→C→E (18%)	B→C→E (0.16)
22.	B→D (50%)	B→D (50%)	B→D (0.45)
23.	B→D→E (30%)	B→D→E (36%)	B→D→E (0.18)
24.	B→E (40%)	B→E (12%)	B→E (0.23)
25.	C (60%)	C (18%)	C (0.52)
26.	C→D (40%)	C→D (48%)	C→D (0.43)
27.	C→D→E (20%)	C→D→E (28%)	C→D→E (0.10)
28.	C→E (30%)	C→E (15%)	C→E (0.16)
29.	D (80%)	D (72%)	D (0.90)
30.	D→E (40%)	D→E (44%)	D→E (0.26)
31.	E (50%)	E (10%)	E (0.32)

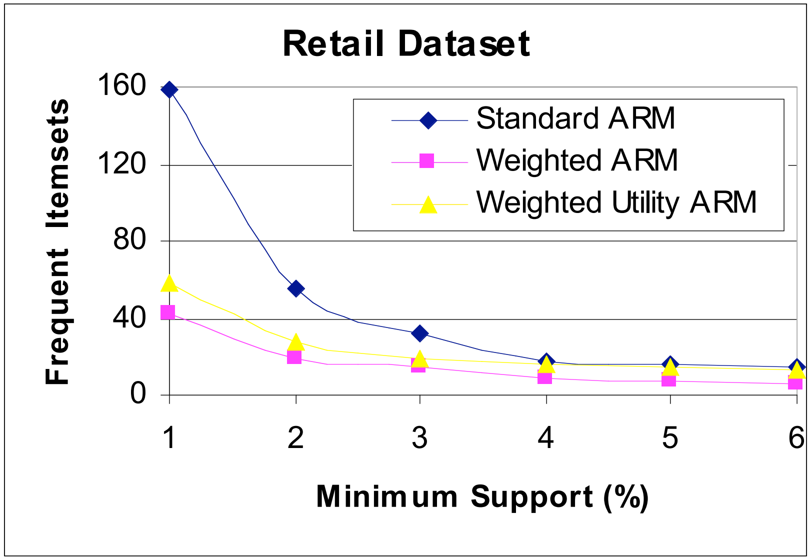
Dataset

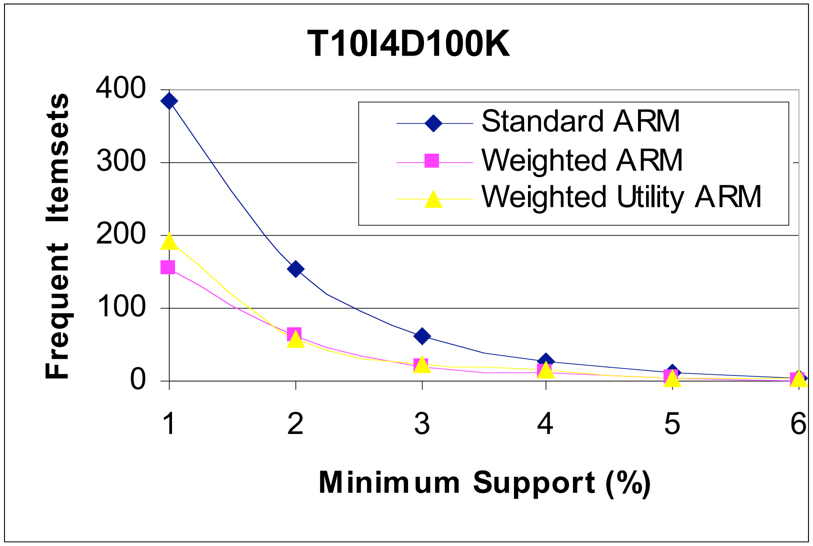
Dataset	No. of Transactions	Distinct Items	Avg. Transaction Size	Max. Transaction Size
Retail	88,162	16,469	10.3	76
T10I4D100K	100,000	1000	10.1	30

- Table characterises the two datasets in terms of
  - number of transactions
  - number of distinct items
  - average transaction size
  - maximum transaction size

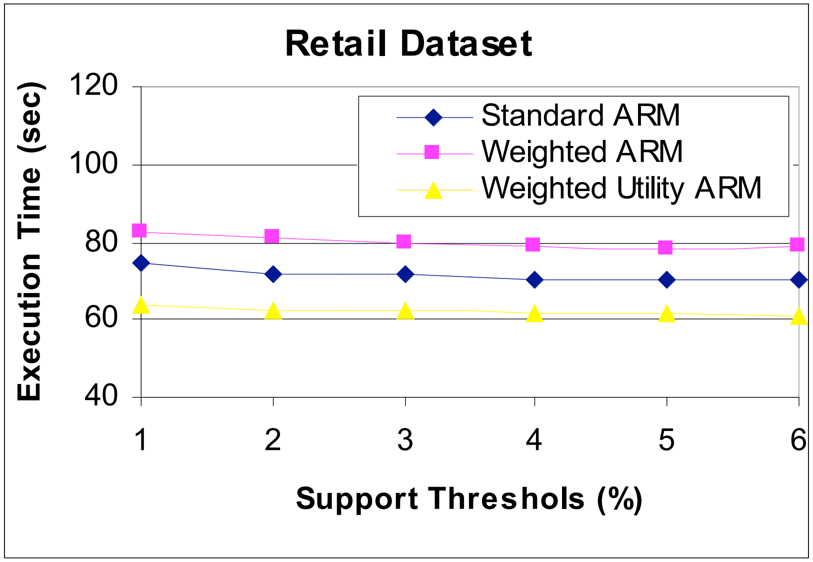
- It is worth mentioning that both datasets contains sparse data, since most association rules discovery algorithms were designed for these types of problems.

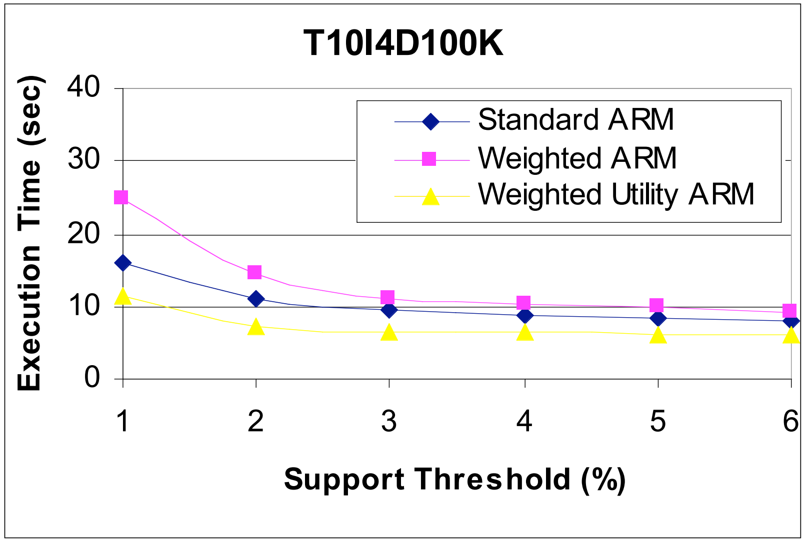
Quality Measures





Performance Measures





Applications

- Proposed approach is widely applicable, e.g.
  - In identifying high profit items with frequent sales, significant weight and high utility, which could be helpful for retail owners and managers to determine
  - valuable items
  - and in decision making for
    - shelf re-arrangements
    - promotional offers
    - catalogue design
    - cross marketing
    - loss leader analysis etc.

Conclusion

- In this paper, we have presented
  - Hybrid framework for mining Weighted Utility ARs
  - Items significance and frequencies
  - Itemsets holds DCP
  - Methodology
  - Experimental evaluation
    - Real and Synthetic datasets
    - Quality Measures
    - Performance Measures
    - WUARM: efficient modified Apriori algorithm
    - The experiments also show that the algorithm is scalable.
  - Application
  - Future work

Liverpool Hope University