



a Bitlab software

Association Rules collaborative tool

Integrated suite for association rule discovering in medical and molecular data

Guided Exercises



Version v1: 8th November 2007.
On-line updated information available at:
<http://chirimoyo.ac.uma.es/arco>

Developed by:
Jesús Jiménez Espada
Javier Ríos
Andrés Rodríguez
Oswaldo Trelles

Report incidences to:
ots@ac.uma.es

Contents

This document contains a Guided Tour through the **ARco** suite and it was created for training purposes with respect to the system options and analysis possibilities. It is not intended for training about the biological / clinical interpretation of the results.

Example data sets can be obtained in **ARco** format directly from the home page at:

<http://chirimoyo.ac.uma.es/arco>

Note: this file contains a complete demo over a simple set of data. ARco is a product evolving continuously and the last updated version is available at the web-site, where new functionalities are announced

Please, submit any recommendation or suggestions from the **ARco**-home page, or directly to ots@ac.uma.es

Contents:

Exercise 1: Extended market basket analysis

Section 1.1: Metadata - Metadata transactions

Section 1.2: Metadata - Experiment transactions

Section 1.3: L - Metadata transactions

Exercise 2: Gene-expression data analysis

Section 1.1: ... transactions

Exercise 1. Extended market basket analysis

This is an extended analysis of the simple market basket exercise used in the introduction section. The same marketDemo100.xls file will be used. Let's have a look of this file

	A	B	C	D	E	F	G	H	I
1				Supplier	A1	A2	A1	A1	A2
2				PriceRange	P1	P1	P2	P2	P1
3				Restrictions	Yes	Not	Yes	Yes	Yes
4	TID	ZIP code	Level(ABCD)	Civilstage	beer	juice	champagne	High Q wine	wine X
5	1	29003	A	Single	4	1			3
6	2	29003	A	Single	5	1			2
7	3	29003	A	Single	7	5			4
8	4	29003	A	Single	3	4			1
9	5	29003	B	Single	2	3			5
10	6	29003	B	Single	1	6	2	1	
11	16	29003	C	Married		2			2
12	17	29003	C	Married		1			1
13	18	29003	B	Married		2	2	2	

As it has been commented this file contains a set of individual transactions from a hypothetical retail store in which five different items are offered. The company also have some information about the customers annotated in the columns "ZIP code {29001, 29002, 29003}", "(economical) Level {A, B, C, D}" and "Civil Stage {Single, Married}". There are also available some information about the items, such as "Supplier {A1, A2}; "Price Range {P1, P2}, and "(alcohol) Restrictions {Yes, Not}

In the table, each cell contains the number of items purchased in each transactions (an empty cell means "No purchase")

Although in this exercise we are not going to use the quantities, observe that it is also possible to focus the attention only in those items that acquired in quantities greater than a given threshold.

It is also noteworthy to comment that Arco has been designed to work especially with gene-expression data, however this market-basket file has a similar structure, and we will use it as an straightforward example for introducing some of the most powerful features of this software.

Starting: (for details on the load procedure see the Introduction or User Manual documents)

- 0) Launch the suite by double click in Arco icon
- 1) Upload the data set. Click "Load" button and browse the file system to locate the appropriated file. Engine; MS-Excel and text-tabulated files are accepted (see File Format annexes)
- 2) The full ARco screenshot is displayed.

Section 1.1. Metadata - Metadata transactions

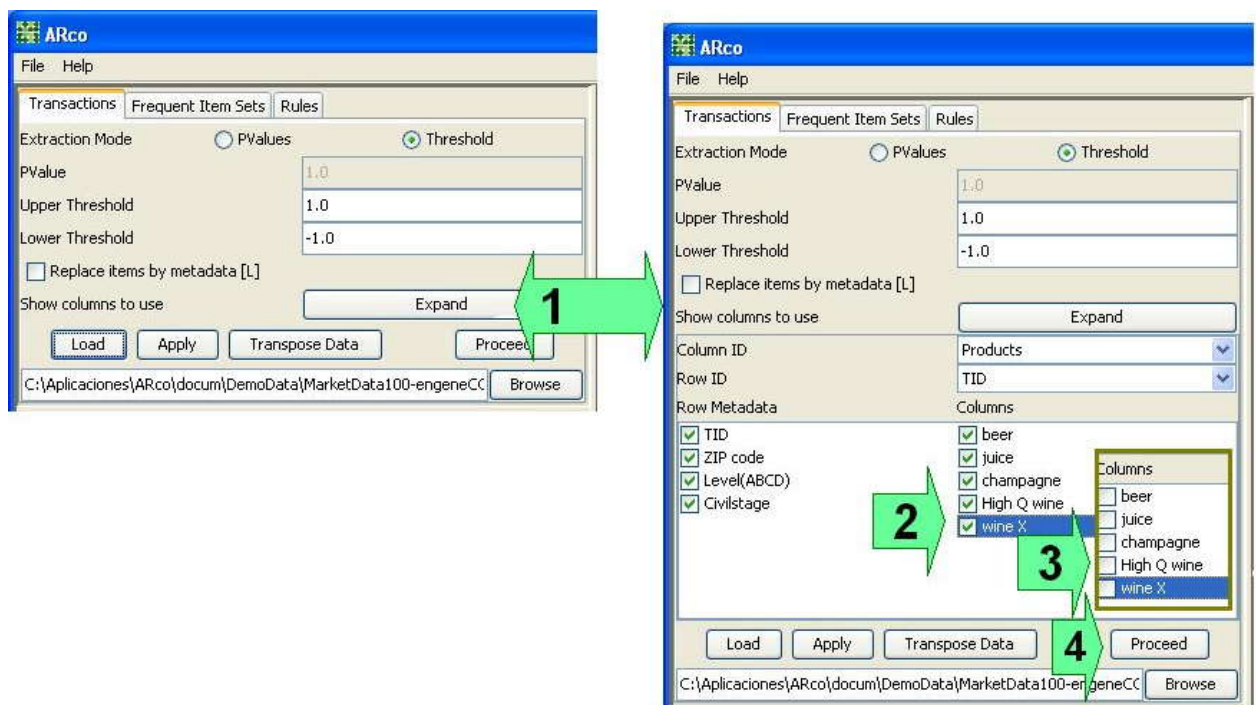
Building transactions with row's metadata

A rule with the following structure: Metadata \Rightarrow Metadata can discover interesting co-occurrences than can lead to the discovering of new knowledge. For instance, the correlation of biological pathways annotations against functional properties could allow to consolidate biological annotations protein databases, but also propose putative annotations to be confirmed. This could be a typical scenario for M \Rightarrow M rules.

Let's use the basket dataset to illustrate the procedure. In our purchase data base we will try to find relationships among customer's metadata in the form of Metadata \Rightarrow Metadata rules.

First select the columns over which the transactions will be built (customer's metadata)

1. Click on "expand" button to show the configuration panel. All the row-metadata and columns will be shown (see figure, step 1).



2. Un-select Columns values (see figure, step 2). These column values are equivalent to experiment values or product bought in this example.
3. Set TID as row (transaction) identifier.
4. Proceed to generate transactions (see figure step 4) (A *.tr file will be generated, set the filename in the "browse" textbox or use the default value. (Results in the Information box: "100 Transactions generated").

Mining frequent itemsets from transaction file

Setting parameters in the “Frequent itemsets tab”

- Set Borglet algorithm; Absolute support type and Unique Mode, and minimal and maximal number of items with the defaults values (see step 1 in the picture)
Notes:
 - support mode is only used when single support is desired;
 - there are two support types: unique, when the same support is used for all the items and multiple if there are different supports.
- Check multiple supports
- Run. Results in the Information box: “39 frequent itemsets generated”.

The image shows three screenshots of the ARco software interface. The first screenshot shows the 'Frequent Item Sets' tab with parameters: Algorithm (Borglet), Support Type (Absolute), Support Mode (Unique), Minimal number of items (1), Maximal number of items (5), Minimal support (5.0), and Maximal support (100.0). The second screenshot shows the 'Rules' tab with Confidence (30.0), Improvement (1.0), and Minimal consequent size (1). The third screenshot shows the 'Rules' tab with Confidence (30.0), Improvement (1.0), Minimal consequent size (1), and various other parameters. Below these are two screenshots of the 'Data View' tab. The first shows a table of frequent itemsets with columns: Antecedent, Consequent, Confidence, Support, ABS Su..., Cover..., Improv..., Lever..., Conv..., Entropy, and RuleID. The second shows a heatmap visualization of the data.

Antecedent	Consequent	Confide...	Support	ABS Su...	Cover...	Improv...	Lever...	Conv...	Entropy	RuleID
[ZIP code:29003]	[Level(ABCD).A]	36,00	9,00	9,00	25,00	3,27	6,25	139,06	0	0
[ZIP code:29001]	[Level(ABCD).D]	50,00	16,00	16,00	32,00	2,08	8,32	152,00	0	1
[ZIP code:29002]	[Level(ABCD).C]	62,79	27,00	27,00	43,00	1,43	8,08	150,50	0	2
[ZIP code:29003]	[Level(ABCD).B]	52,00	13,00	13,00	25,00	2,48	7,75	164,58	0	3
[ZIP code:29003]	[Civilstage.Married]	48,00	12,00	12,00	25,00	1,00	0,00	100,00	0	4
[ZIP code:29003]	[Civilstage.Single]	52,00	13,00	13,00	25,00	1,00	0,00	100,00	0	5
[ZIP code:29001]	[Civilstage.Married]	53,13	17,00	17,00	32,00	1,11	1,64	110,93	0	6
[ZIP code:29002]	[Civilstage.Single]	55,81	24,00	24,00	43,00	1,07	1,64	108,63	0	7

Rule production

Once we have the frequent itemset collection, next step is rule production.

Go to the “Rules” tab and click on the Run button. Results in the Information box: “39 rules generated”. (See section 2 in the picture)

In this case we have used the “Both” default value for items location in the rule. However, in some cases, questions such as “which is the profile of customer by “zip code”. Observe that this information can help in designing a mailing campaign in the supermarket with specific products for specific zones. To get this result we can modify the setting by using the “Expand” button (see figure, section 3). In this case we want to have sip code in the consequent and Economical level & Civil stage in the consequent.

Launch again the “Run” button in the “Rules” tab to obtain 8 rules in this example (see figure, section 4).

Click in the first rule and the transactions that hold the rule are displayed and items in the rule are highlighted (see section 5) in figure; and a graphical representation of the rule is shown in the frame 6 in the picture.

Some rules among metadata arise to show relations like: particular postal code (ZIP) implies an concrete economic level.

Section 1.2. Metadata - Experiment transactions

One of the most interesting associations can be derived from Metadata / Experiments rules. Following the market-basket example, a “zip-code ⇒ products” association rule can provide information to be used in several activities.

- 1) Check the appropriated items are selected to produce transactions and proceed (see section 1 in the picture): 100 transactions will be generated.
- 2) Produce frequent itemsets with Borglet algorithm / Unique Mode and Absolute Type (30 itemsets are obtained)
- 3) Set the details of the rule in the rule’s tab (see section 2 in the picture). 4 rules are produced (see section 3 in the picture).

The screenshot shows the configuration of an experiment in a data mining software. It is divided into three main sections:

Section 1: Transaction Extraction

- Extraction Mode: PValues, Threshold
- PValue: 1.0
- Upper Threshold: 1.0
- Lower Threshold: -1.0
- Replace items by metadata [L]
- Show columns to use: Expand
- Column ID: Products
- Row ID: TID
- Row Metadata:
 - TID
 - ZIP code
 - Level(ABCD)
 - Civilstage
- Columns:
 - beer
 - juice
 - champagne
 - High Q wine
 - wine X

Section 2: Rule Configuration

- Confidence: 30.0
- Improvement: 1.0
- Minimal consequent size: 1
- Appearance: Expand
- ZIP code: Ant., Con., Both, None
- [+beer]: Ant., Con., Both, None
- [+juice]: Ant., Con., Both, None
- [+champagne]: Ant., Con., Both, None
- [+High Q wine]: Ant., Con., Both, None
- [+wine X]: Ant., Con., Both, None
- [-beer]: Ant., Con., Both, None
- [-juice]: Ant., Con., Both, None
- [-champagne]: Ant., Con., Both, None
- [-High Q wine]: Ant., Con., Both, None
- [-wine X]: Ant., Con., Both, None

Section 3: Rules Table

Antecedent	Consequent	Confidence	Support	ABS Sup...	Coverage	Improve...	Leverage	Conviction	Entropy	RuleID
[ZIP code.29003]	[+beer]	44,00	11,00	11,00	25,00	2,93	7,25	151,79	0	0
[ZIP code.29003]	[+juice]	60,00	15,00	15,00	25,00	1,43	4,50	145,00	0	1
[ZIP code.29001]	[+champag...]	53,13	17,00	17,00	32,00	1,77	7,40	149,33	0	2
[ZIP code.29001]	[+High Q w...]	50,00	16,00	16,00	32,00	1,25	3,20	120,00	0	3

Section 1.3. L - Metadata transactions

One of the most interesting features of ARco is the possibility to correlate row-metadata with column-metadata through the items that are bought by customers.

To use this outstanding ability of ARco proceed by Loading (1) the data file; (2) set On the “Replace by metadata [L]” check box and click on the “Expand” button (3) to display all the data labels. Set on the fields (4) and proceed (5). As result 100 transactions will be generated.

The screenshot illustrates the ARco software interface for generating transactions from metadata. The main window is titled 'Transactions' and has three tabs: 'Transactions', 'Frequent Item Sets', and 'Rules'. The 'Transactions' tab is active, showing the following settings:

- Extraction Mode:** Threshold (selected), PValues (unselected)
- PValue:** 0.1
- Upper Threshold:** 0.0
- Lower Threshold:** 0.0
- Replace items by metadata [L]:** Checked
- Show columns to use:** Expand button
- Column ID:** Exp
- Row ID:** TID
- Row metadata:** ZIP code (checked)
- Columns:** beer, juice, champagne, High Q wine, wine X (all checked)
- Column Metadata:** Supplier, PriceRange, Restrictions (all checked)

Buttons at the bottom include 'Load', 'Apply', 'Transpose Data', and 'Proceed'. A 'Browse' button is also present. The 'Transaction info' window shows the following statistics:

- Longest Transaction: 7
- Shortest Transaction: 4
- Average Length: 6
- Number of transactions: 100
- Longest Transaction: 7
- Shortest Transaction: 4
- Average Length: 6

The summary window at the bottom shows the generated transactions with their metadata:

```

Transactions
Experiments used: [beer, juice, champagne, High Q wine, wine X]
Metadata used: [ZIP code]
Experiments metadata used: [Supplier, PriceRange, Restrictions]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]
[ZIP code.29003] [Supplier.A1] [Supplier.A2] [PriceRange.P2] [PriceRange.P1] [Restrictions.Yes] [Restrictions.Not]

```