Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Παράδειγμα – Κανόνων Συσχετίσεων

Πρόβλημα

Η Μαρία είναι διευθύντρια σε ένα σχολικό συγκρότημα το οποίο έχει πολλές και μεγάλες ανάγκες που δυστυχώς όμως δεν υπάρχουν οι απαραίτητοι πόροι για να ικανοποιηθούν.

Γνωρίζει όμως ότι πολλοί γονείς δραστηριοποιούνται σε διάφορους συλλόγους και οργανώσεις και θεωρεί πως αν καταφέρει να χρησιμοποιήσει τις γνωριμίες που έχουν οι γονείς ή πολύ περισσότερο αν καταφέρει να κάνει γονείς που ανήκουν σε διαφορετικούς συλλόγους να συνεργαστούν μεταξύ τους, τα οφέλη για το σχολικό συγκρότημα θα είναι πολύ μεγάλα.

Πρόβλημα

Οπότε αυτό που θα πρέπει να κάνει είναι να συλλέξει κάποια στοιχεία σχετικά με τους γονείς και τους συλλόγους στους οποίους συμμετέχουν και στη συνέχεια να διερευνήσει ενδεχόμενες συνδέσεις μεταξύ των συλλόγων.

Αφού συγκεντρώσει τα πρώτα αποτελέσματα τότε θα μπορέσει να έρθει σε επαφή με τους γονείς και να τους ζητήσει να χρησιμοποιήσουν τις γνωριμίες τους προκειμένου να ξεκινήσει η υλοποίηση κάποιων πολύ σημαντικών έργων για το σχολείο τα οποία όμως είχαν καθυστερήσει για πολλά χρόνια.

(M. North, Data Mining for the Masses, 2012

Διασκευή παραδείγματος κεφαλαίου 5)

Survey & RapidMiner

Από κάθε σύλλογο κ.λπ. ζητήθηκε να συμμετέχει σε μία έρευνα και μάλιστα ο επικεφαλής λαμβάνει μοναδικό συνθηματικό (unique password) που μοιράζει στα μέλη του συλλόγου του. Τα μέλη ενθαρρύνονται να συμπληρώσουν στοιχεία - απαντήσεις σε ειδική ιστοσελίδα που δημιουργήθηκε. Όταν ένα φυσικό πρόσωπο συμπληρώνει τις απαντήσεις του καταγράφεται και το συνθηματικό του συλλόγου του για να γνωρίζουμε πόσα άτομα συμμετέχουν στην έρευνα από κάθε σύλλογο. Η έρευνα κρατά ένα μήνα και να τα attributes του data set της έρευνας.

Attributes

Elapsed_Time: Χρόνος συμπλήρωσης πχ 4.5 δηλαδή 4 λεπτά και 30 δευτερόλεπτα

Time_in_Community: Πόσα χρόνια ζει το άτομο στο Δήμο (Short: 0-2 years, Medium: 3-9 years, Long: 10+ years)

Gender:

Working: yes/no

Age:

Family: yes/no, σημειώνει αν το άτομο συμμετέχει σε σύλλογους όπως σύλλογος κηδεμόνων, σύλλογος παιδικής αναψυχής, σύλλογος διοργάνωσης παιδικών πρωταθλημάτων.

Hobbies: yes/no, σημειώνει αν συμμετέχει σε σύλλογους όπως σύλλογος ραδιοερασιτεχνών, σύλλογος ποδηλατών

Social_Club: yes/no, σημειώνει αν συμμετέχει σε σύλλογους με κάποια κοινωνική δραστηριότητα όπως φιλανθρωπικός σύλλογος κυριών

Political: yes/no, σημειώνει αν συμμετέχει σε πολιτικό κόμμα

Professional: yes/no, σημειώνει αν συμμετέχει σε επαγγελματικούς σύλλογους

Religious: yes/no, σημειώνει αν συμμετέχει σε κάποια εκκλησία πχ βουδιστές

Support_Group: yes/no, σημειώνει αν συμμετέχει σε σύλλογους στήριξης όπως σύλλογος στήριξης ανώνυμων αλκοολικών

Ο πίνακας δεδομένων (Data view) αποτελείται από 3483 observations (examples) και 12 regular attributes.

Row_no	Elapsed_Time	Time_in_Community	Gender	Working	Age	Family	Hobbies	Social_Club	Political	Professional	Religious	Support_Group
1	8.71	Short	M	No	53.0	1	0	0	0	0	0	0
2	5.24	Medium	F	No	31.0	0	0	0	0	0	1	1
3	4.22	Medium	M	No	42.0	1	1	0	0	1	0	0
4	4.81	Long	F	No	30.0	0	0	0	0	0	0	0
5	3.95	Long	M	Yes	29.0	0	0	0	1	1	0	1
6	9.35	Long	F	No	40.0	0	0	0	0	1	0	0
7	2.91	Medium	F	Yes	33.0	0	0	0	0	0	0	1
8	4.54	Medium	M	Yes	27.0	1	1	1	0	0	1	0
9	4.79	Short	F	No	50.0	1	1	0	0	1	1	0
10	3.07	Medium	M	No	28.0	0	0	0	0	0	1	1
11	2.2	Medium	F	No	20.0	1	1	0	0	1	0	0
12	2.77	Medium	F	Yes	54.0	1	0	1	0	0	1	0
13	7.32	Long	M	Yes	48.0	1	0	0	0	0	0	0
14	2.23	Short	F	Yes	48.0	0	0	0	0	1	0	0
15	2.13	Medium	F	No	25.0	1	1	1	0	0	1	0
16	7.15	Short	F	Yes	25.0	0	0	0	0	0	0	0
17	8.63	Long	M	Yes	38.0	1	0	0	0	1	1	0
18	3.77	Medium	F	Yes	34.0	0	0	0	0	0	0	0
19	2.2	Medium	F	Yes	32.0	0	0	0	0	0	1	0
20	6.52	Long	F	Yes	43.0	1	0	0	0	0	0	1
21	6.42	Medium	M	Yes	43.0	1	1	0	0	0	0	1
22	7.4	Medium	F	Yes	49.0	0	1	0	0	1	1	0
23	6.33	Short	F	Yes	43.0	0	1	0	0	1	1	0
24	4.45	Medium	M	Yes	49.0	1	0	1	0	1	0	0
25	6.1	Medium	M	Yes	35.0	1	1	0	0	0	0	0

στατιστική του συνόλου δεδομένων (Statistics (Meta Data) View)

Name	Type	Statistics	Range
Elapsed_time	Real	Avg=5.922±2.293	[2.010, 10.150]
Time_in_community	Polynominal	Mode=Long(1485) Least=Short(714)	Short(714) Medium(1304) Long(1406)
Gender	Binominal	Mode=F(1790) Least=M(1693)	M(1693), F(1790)
Working	Binominal	Mode=Yes(1744) Least=No(1739)	No(1739), Yes(1744)
Age	Integer	Avg=36.731±10.647	[17.000, 57000]
Family	Integer	Avg=0.390±0.488	[0.000, 1.000]
Hobbies	Integer	Avg=0.300±0.458	[0.000, 1.000]
Social_club	Integer	Avg=0.188±0.391	[0.000, 1.000]
Political	Integer	Avg=0.094±0.292	[0.000, 1.000]
Professional	Integer	Avg=0.324±0.468	[0.000, 1.000]
Religious	Integer	Avg=0.419±0.493	[0.000, 1.000]
Support_group	Integer	Avg=0.159±0.366	[0.000, 1.000]

https://sites.google.com/site/dataminingforthemasses/

Elapsed_Time,Time_in_Community,Gender,Working,Age,Family,Hobbies,Social_Club,Political, Professional,Religious,Support_Group

8.71,Short,M,No,53,1,0,0,0,0,0,0

5.24,Medium,F,No,31,0,0,0,0,0,1,1

4.22,Medium,M,No,42,1,1,0,0,1,0,0

4.81,Long,F,No,30,0,0,0,0,0,0,0

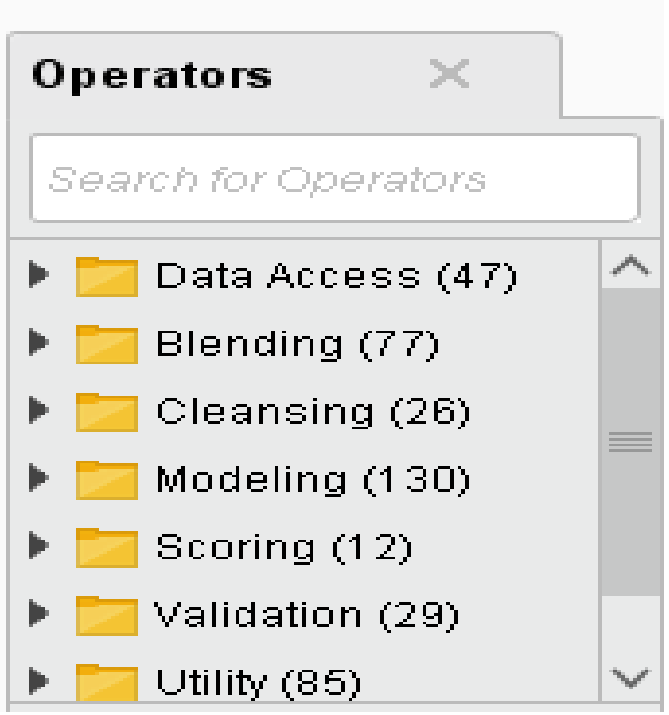
3.95,Long,M,Yes,29,0,0,0,1,1,0,1

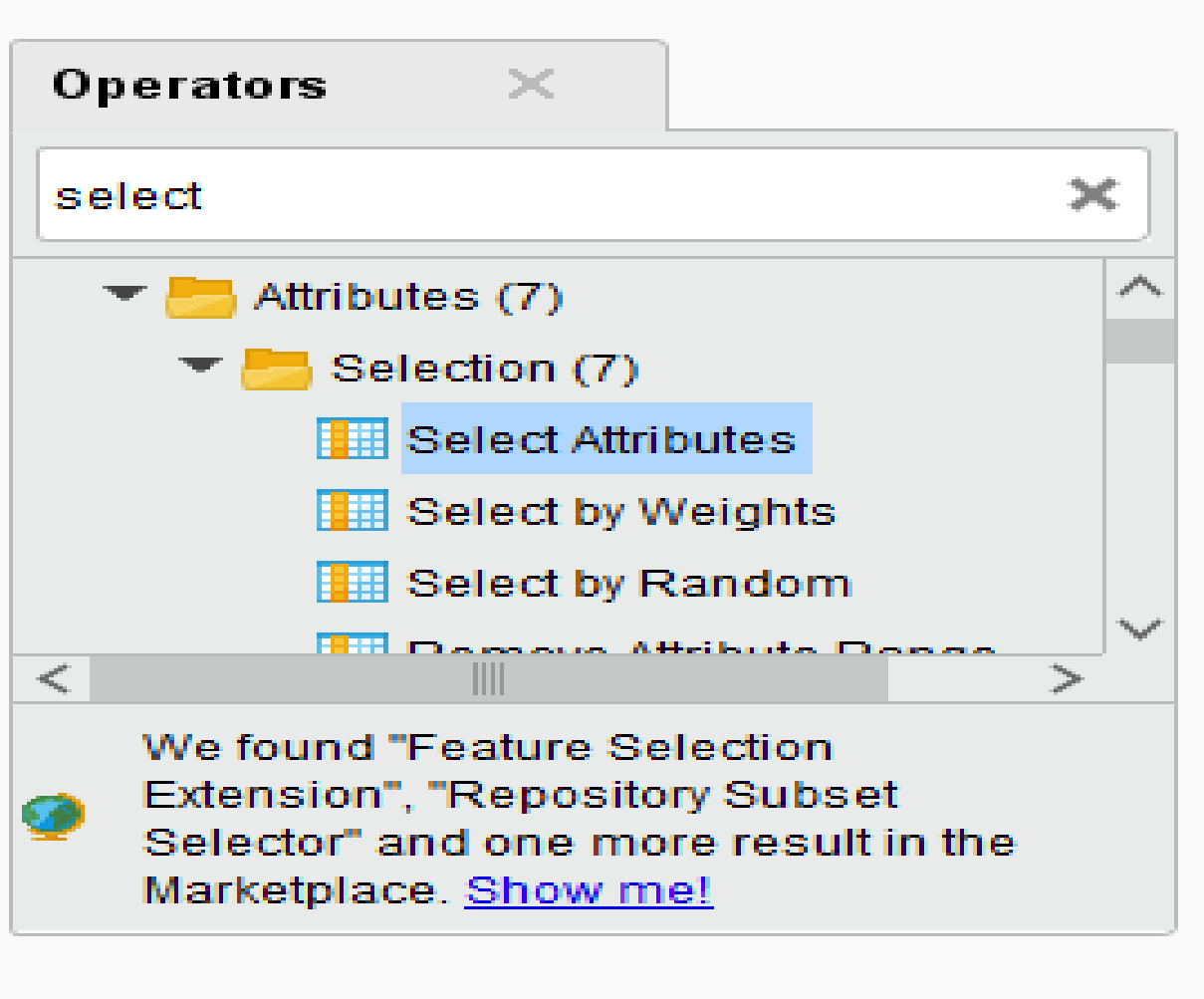
9.35,Long,F,No,40,0,0,0,0,1,0,0

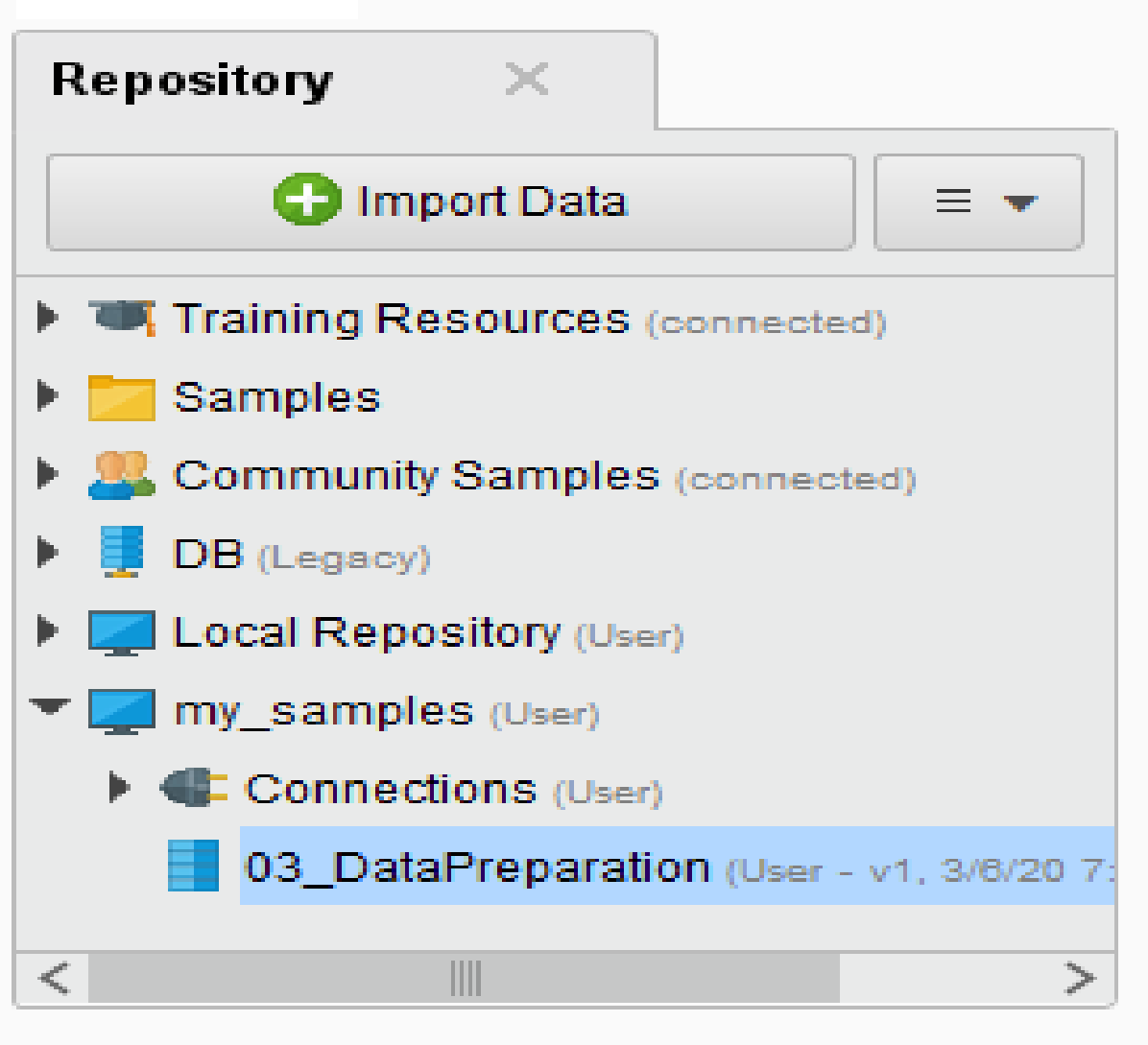
2.91,Medium,F,Yes,33,0,0,0,0,0,0,1

4.54,Medium,M,Yes,27,1,1,1,0,0,1,0

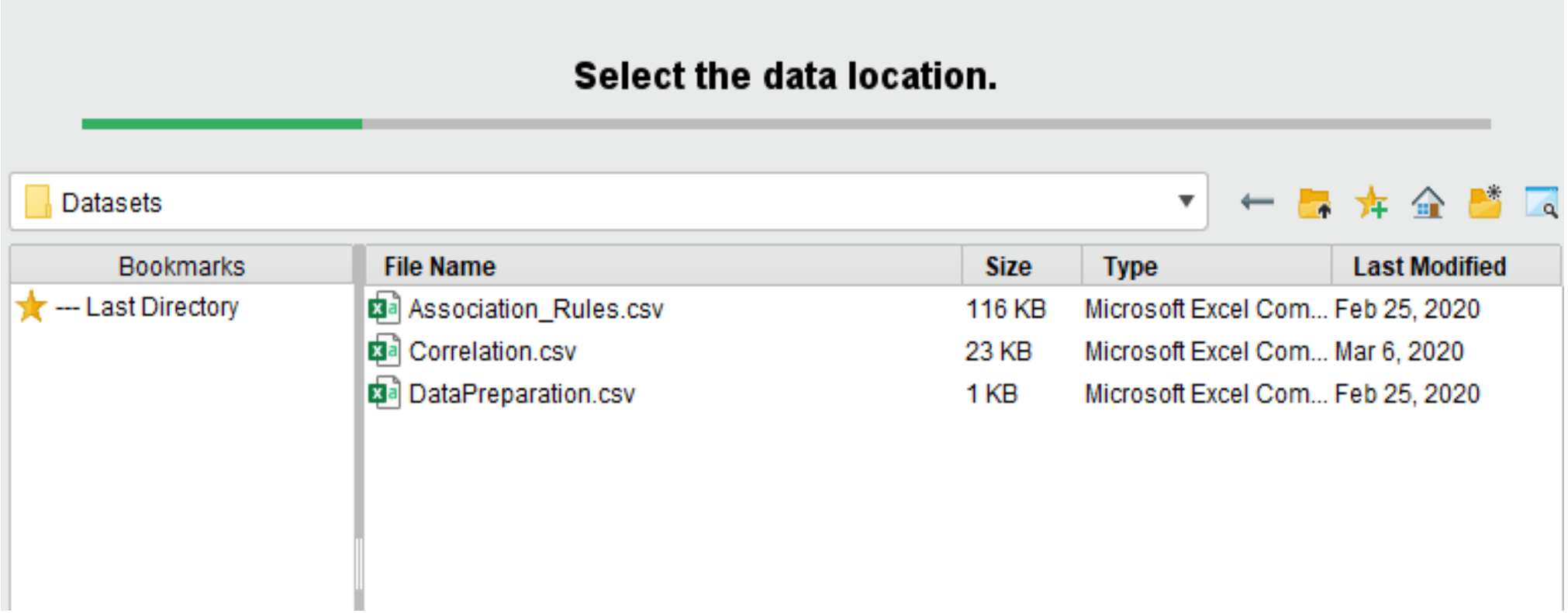
εργαλείο Rapid Miner






Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
DB (Legacy)
Local Repository (User)
my_samples (User)
Connections (User)
03_DataPreparation (User - v1, 3/6/20 7:

Φορτώνουμε το σύνολο δεδομένων Association Rules.csv (Import)



Αρχικά γίνεται η εισαγωγή και η προετοιμασία των δεδομένων (Data preparation) στο εργαλείο RapidMiner.
Εισάγουμε το dataset Association_Rules.csv στο παράθυρο Process Window.
Βλέπουμε δεδομένα και στατιστική
Στη συνέχεια χρησιμοποιούμε operator Select Attribute και επιλέγουμε τα attributes που επιθυμούμε.

Row No.	Family	Hobbies	Social_Club	Political	Professional	Religious	Support_Group
1	1	0	0	0	0	0	0
2	0	0	0	0	0	1	1
3	1	1	0	0	1	0	0
4	0	0	0	0	0	0	0
5	0	0	0	1	1	0	1
6	0	0	0	0	1	0	0
7	0	0	0	0	0	0	1
8	1	1	1	0	0	1	0
9	1	1	0	0	1	1	0
10	0	0	0	0	0	1	1
11	1	1	0	0	1	0	0
12	1	0	1	0	0	1	0
13	1	0	0	0	0	0	0
14	0	0	0	0	1	0	0
15	1	1	1	0	0	1	0

ExampleSet (3,483 examples, 0 special attributes, 7 regular attributes)

Statistics view

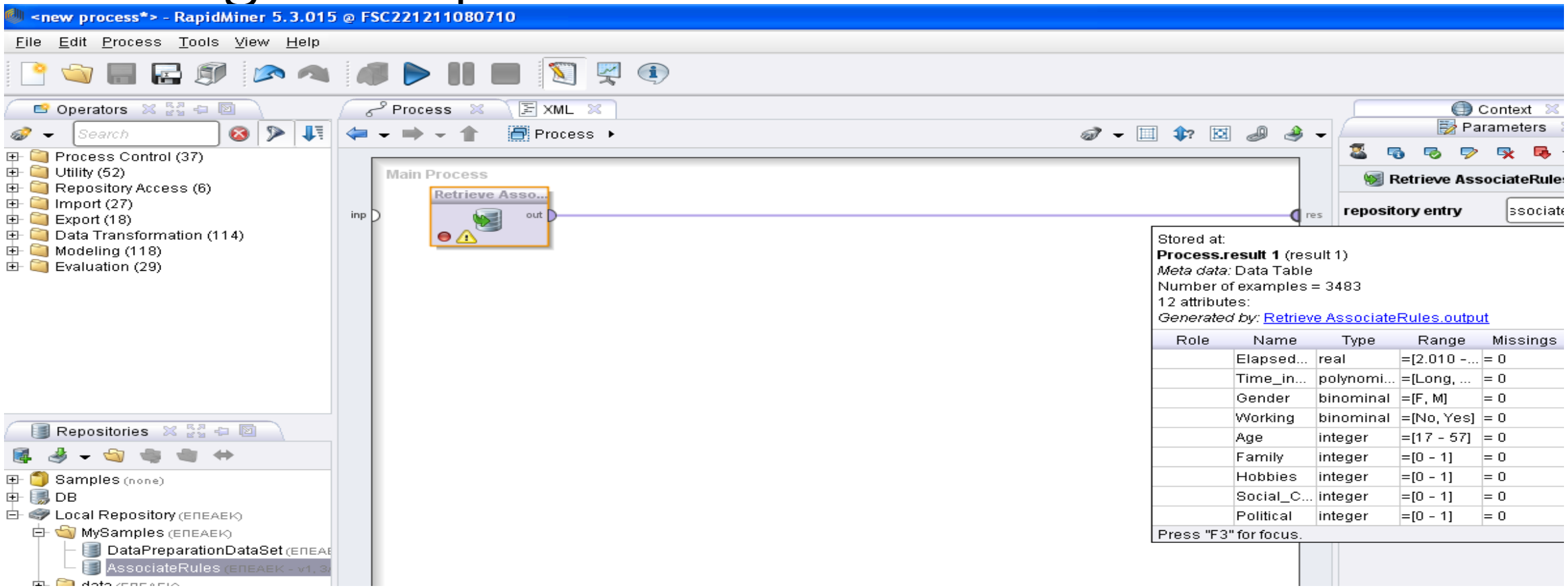
Filter (7 / 7 attributes): Search for Attributes

Name	Type	Missing	Statistics		
			Min	Max	Average
Family	Integer	0	0	1	0.390
Hobbies	Integer	0	0	1	0.300
Social_Club	Integer	0	0	1	0.188
Political	Integer	0	0	1	0.094
Professional	Integer	0	0	1	0.324
Religious	Integer	0	0	1	0.419
Support_Group	Integer	0	0	1	0.159

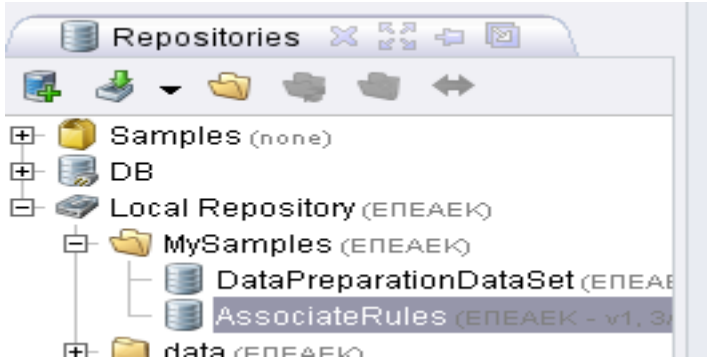
Metadata view

Name	Type	Statistics	Range	Missings
Elapsed_time	Real	Avg=5.922±2.293	[2.010, 10.150]	0
Time_in_community	Polynominal	Mode=Long(1485) Least=Short(714)	Short(714) Medium(1304) Long(1406)	
Gender	Binominal	Mode=F(1790) Least=M(1693)	M(1693), F(1790)	
Working	Binominal	Mode=Yes(1744) Least=No(1739)	No(1739), Yes(1744)	
Age	Integer	Avg=36.731±10.647	[17.000, 57000]	
Family	Integer	Avg=0.390±0.488	[0.000, 1.000]	
Hobbies	Integer	Avg=0.300±0.458	[0.000, 1.000]	
Social_club	Integer	Avg=0.188±0.391	[0.000, 1.000]	
Political	Integer	Avg=0.094±0.292	[0.000, 1.000]	
Professional	Integer	Avg=0.324±0.468	[0.000, 1.000]	
Religious	Integer	Avg=0.419±0.493	[0.000, 1.000]	
Support_group	Integer	Avg=0.159±0.366	[0.000, 1.000]	

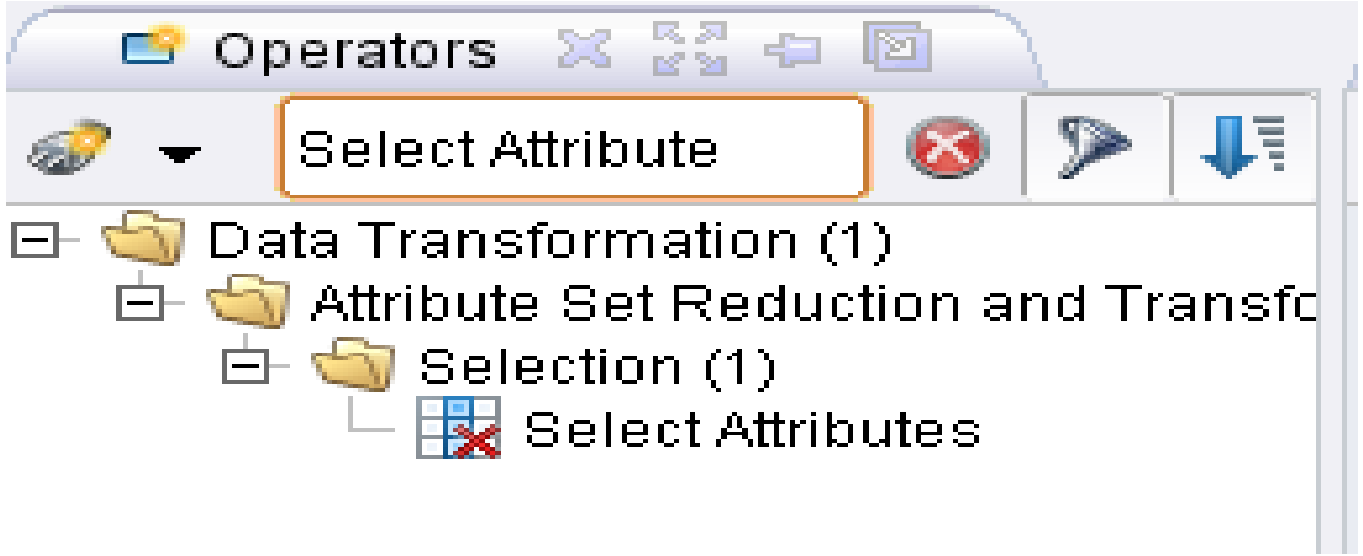
Design Perspective



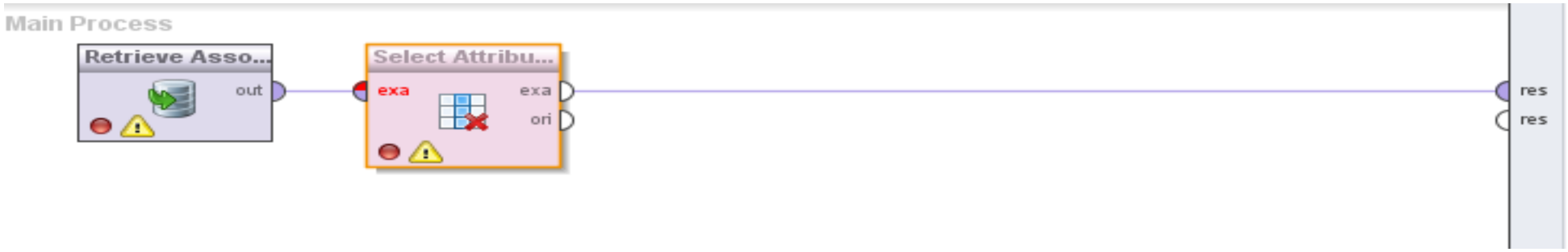
Drag and Drop



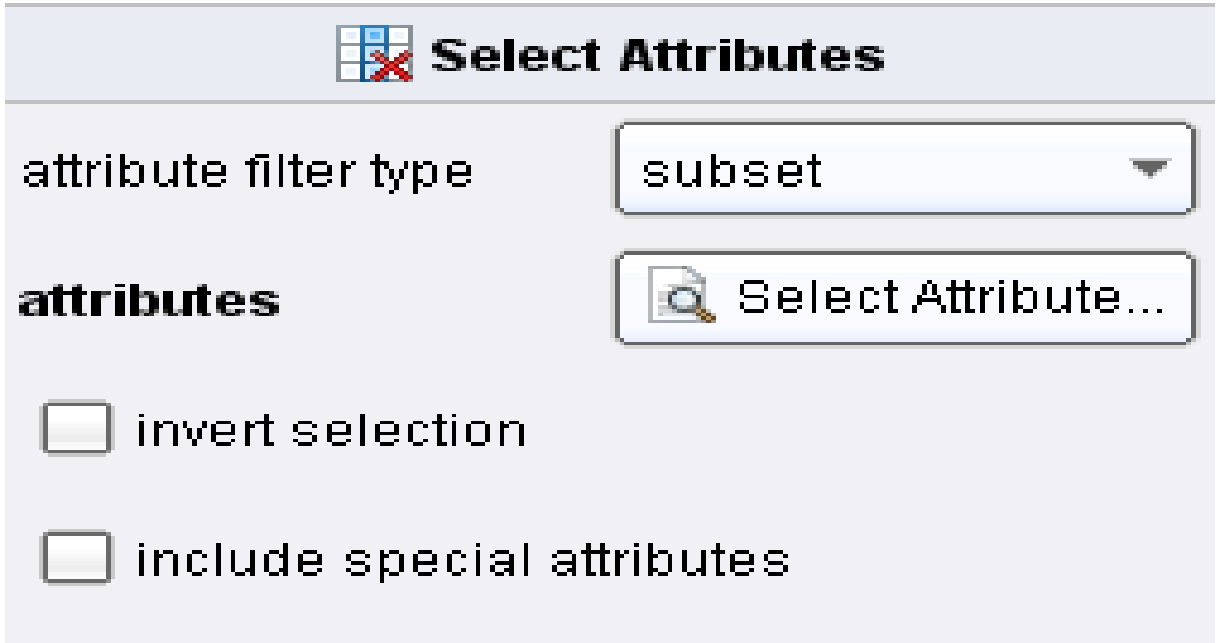
Data preparation: Select Attribute



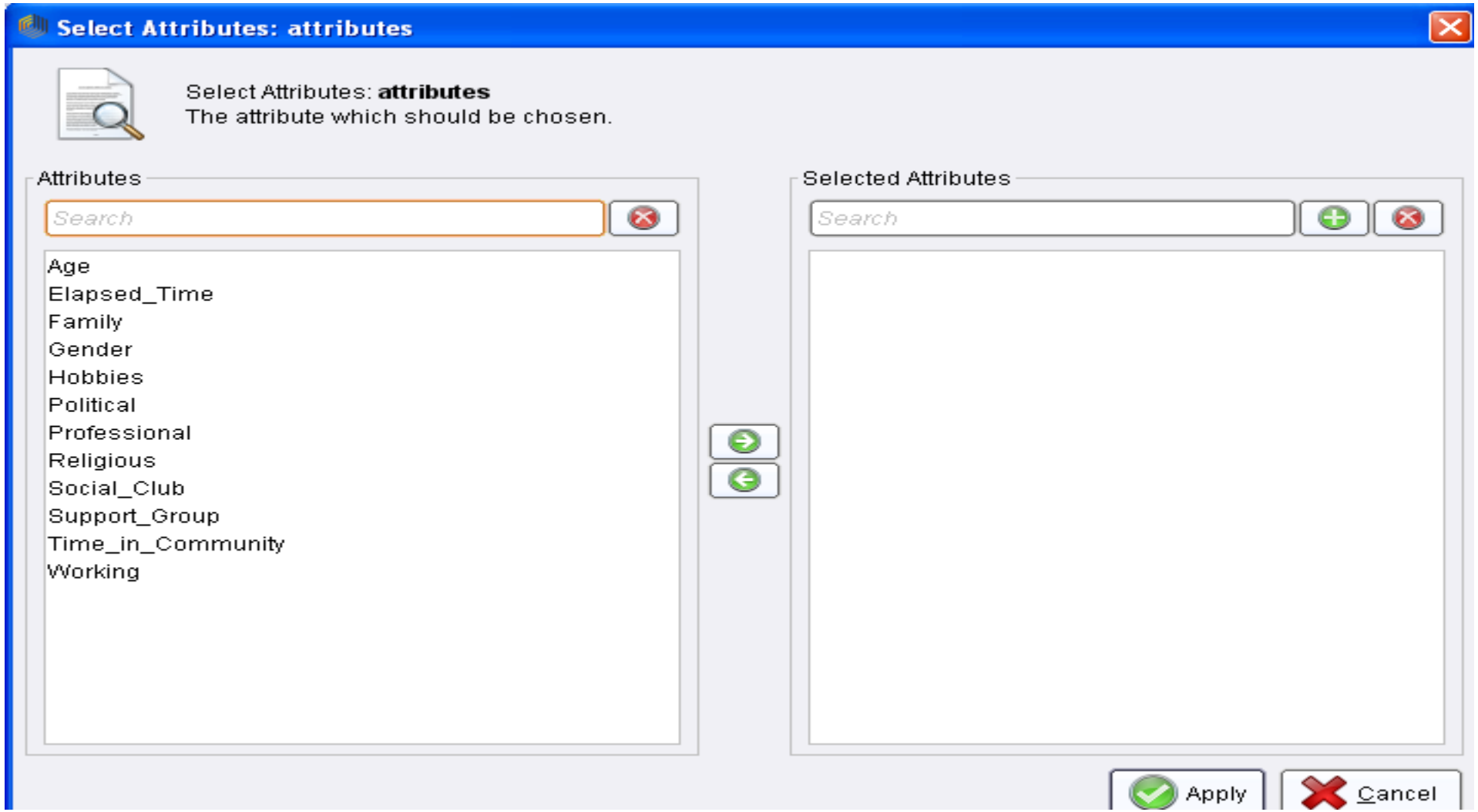
Add select attribute operator to your data mining stream



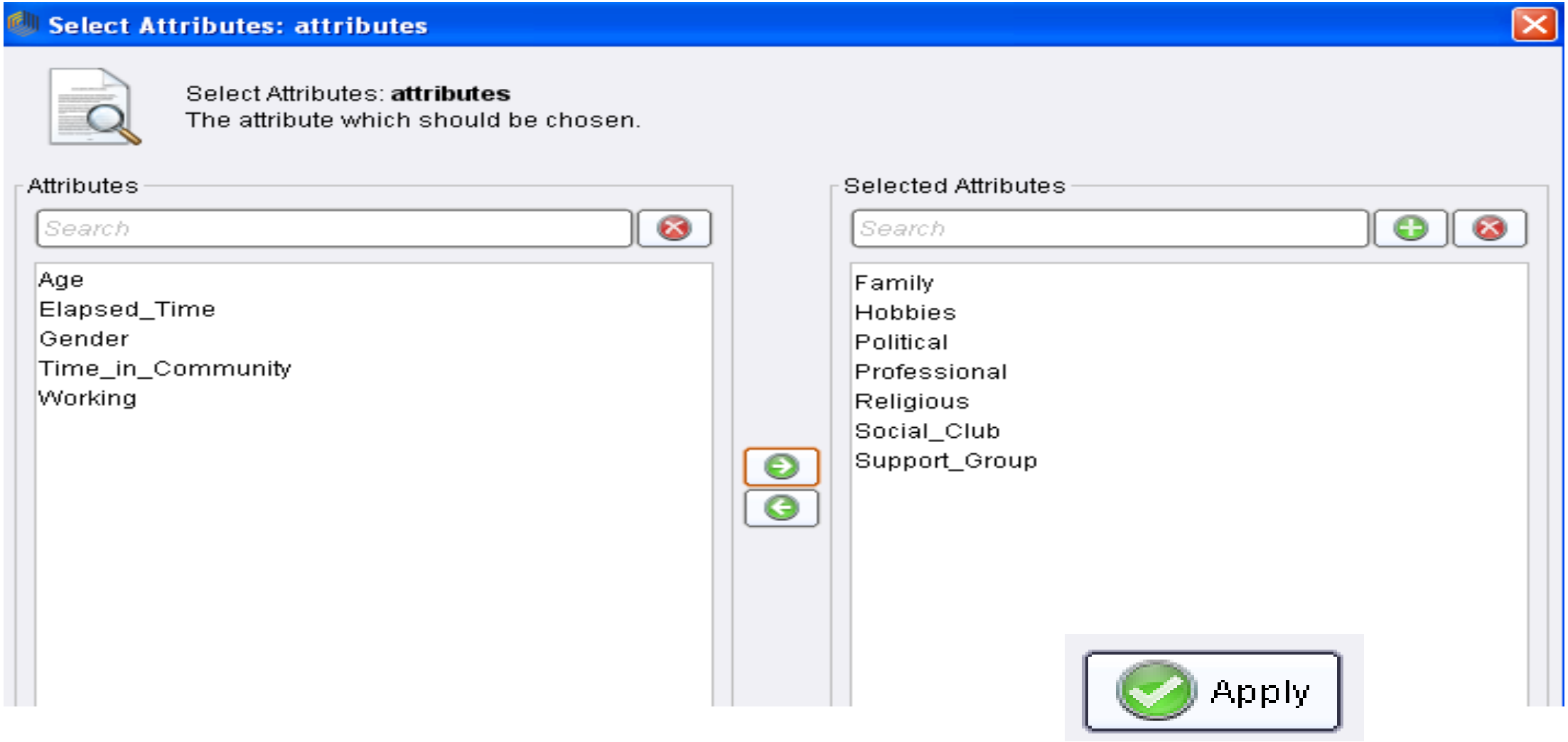
Parameters pane and Select Attributes



Select Attribute...


Select Attributes: attributes
Select Attributes: attributes
The attribute which should be chosen.
Attributes
Search
Age
Elapsed_Time
Family
Gender
Hobbies
Political
Professional
Religious
Social_Club
Support_Group
Time_in_Community
Working
Selected Attributes
Search
Apply
Cancel

Selected attributes for inclusion



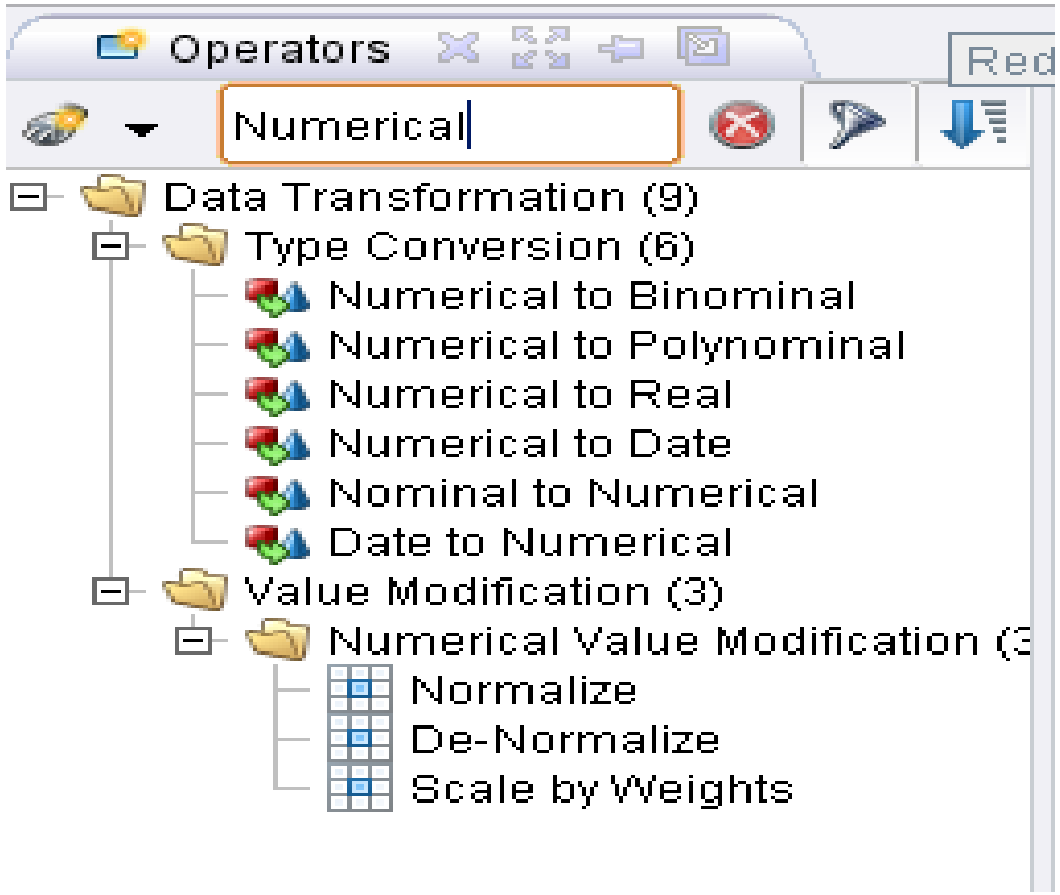
operator **Numerical to Binominal**

- Στο εργαλείο RapidMiner χρησιμοποιείται ο τύπος δεδομένων binominal. Ο τύπος δεδομένων Binominal σημαίνει ότι ένα Attribute λαμβάνει μία από δύο τιμές που μπορεί να είναι αριθμητικές (numeric) ή χαρακτήρες (character), π.χ. true/false.
- Προσοχή! Ο γνωστός τύπος δεδομένων Binomial σημαίνει ότι ένα Attribute λαμβάνει μία από δύο τιμές που μπορεί να είναι αριθμητικές, συνήθως 0 και 1.

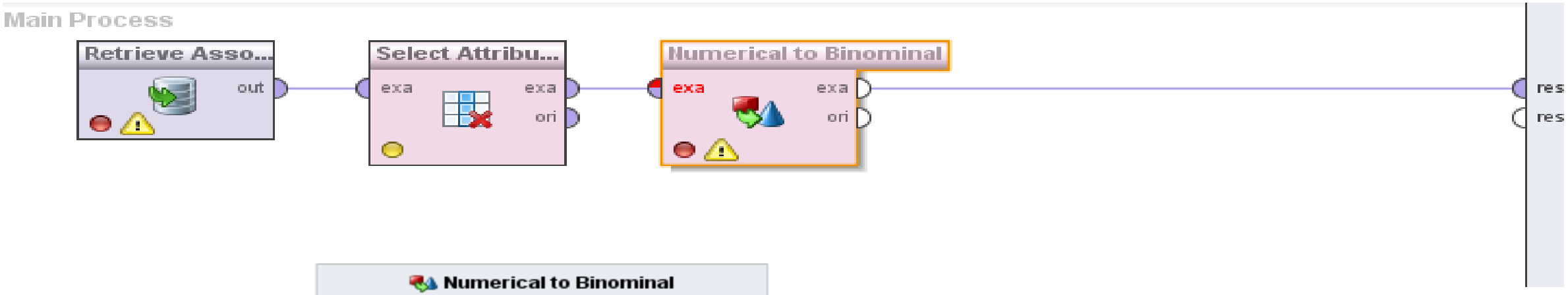
operator **Numerical to Binominal**

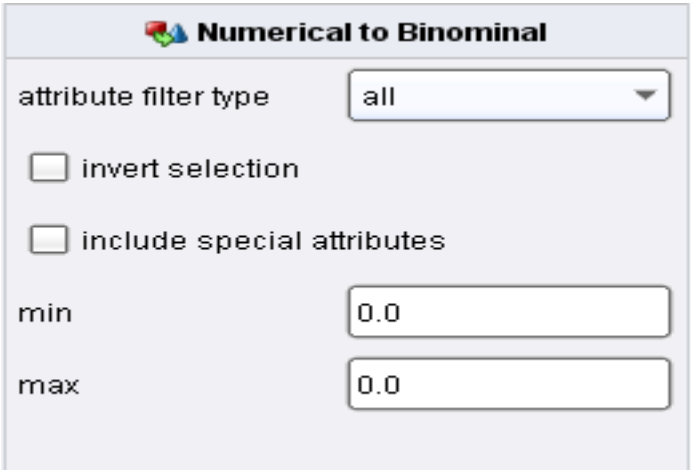
- Χρησιμοποιούμε operator **Numerical to Binominal** γιατί ο μετασχηματισμός του τύπου δεδομένων από **Numerical** (0/1) σε **Binominal** (TRUE/FALSE) αποτελεί προϋπόθεση για την εξαγωγή Association Rules

Data preparation: Numerical to Binominal



Drag the Numerical to Binominal operator into your stream.



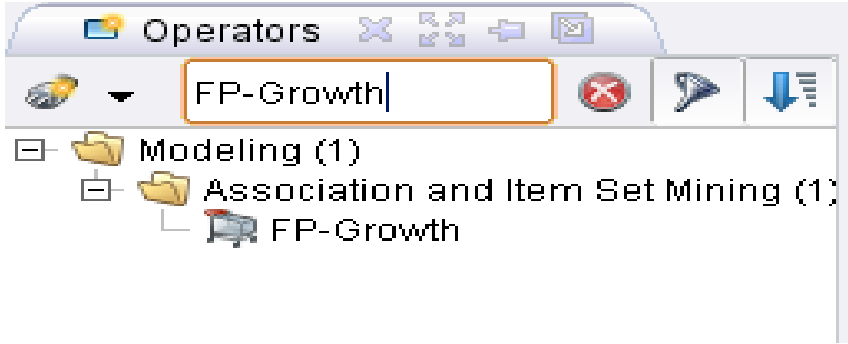


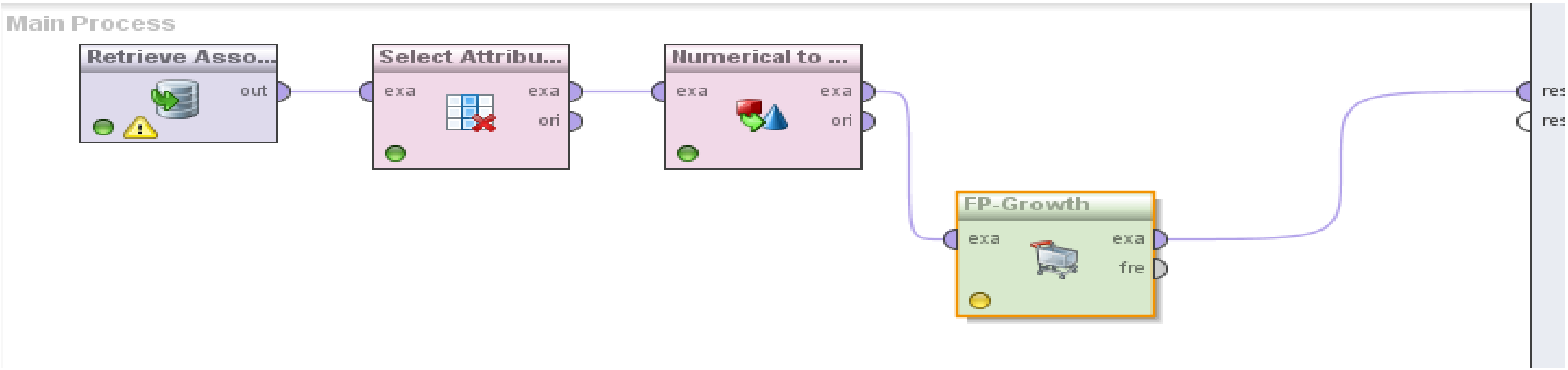
Row_no	Family	Hobbies	Social_club	Political	Professional	Religious	Support_group
1	true	false	False	false	false	false	false
2	false	false	False	false	false	true	true
3	true	true	False	false	true	false	false
4	false	false	False	false	false	false	false
5	false	false	False	true	true	false	true
6	false	false	False	false	true	false	false
7	false	false	False	false	false	false	true
8	true	true	True	false	false	true	false
9	true	true	False	false	true	true	false
10	false	false	False	false	false	true	true
11	true	true	False	false	true	false	false
12	true	false	True	false	false	true	false
13	true	false	False	false	false	false	false
14	false	false	False	false	true	false	false
15	true	True	True	false	false	true	false
16	false	False	False	false	false	false	false
17	true	False	False	false	true	true	false
18	false	False	False	false	false	false	false
19	false	False	False	false	false	true	false
20	true	False	False	false	false	false	true
21	true	True	False	false	false	false	true
22	false	True	False	false	true	true	false
23	false	True	False	false	true	true	false
24	true	False	True	false	true	false	false

Meta Data View Example set 3483examples, 7regular attributes

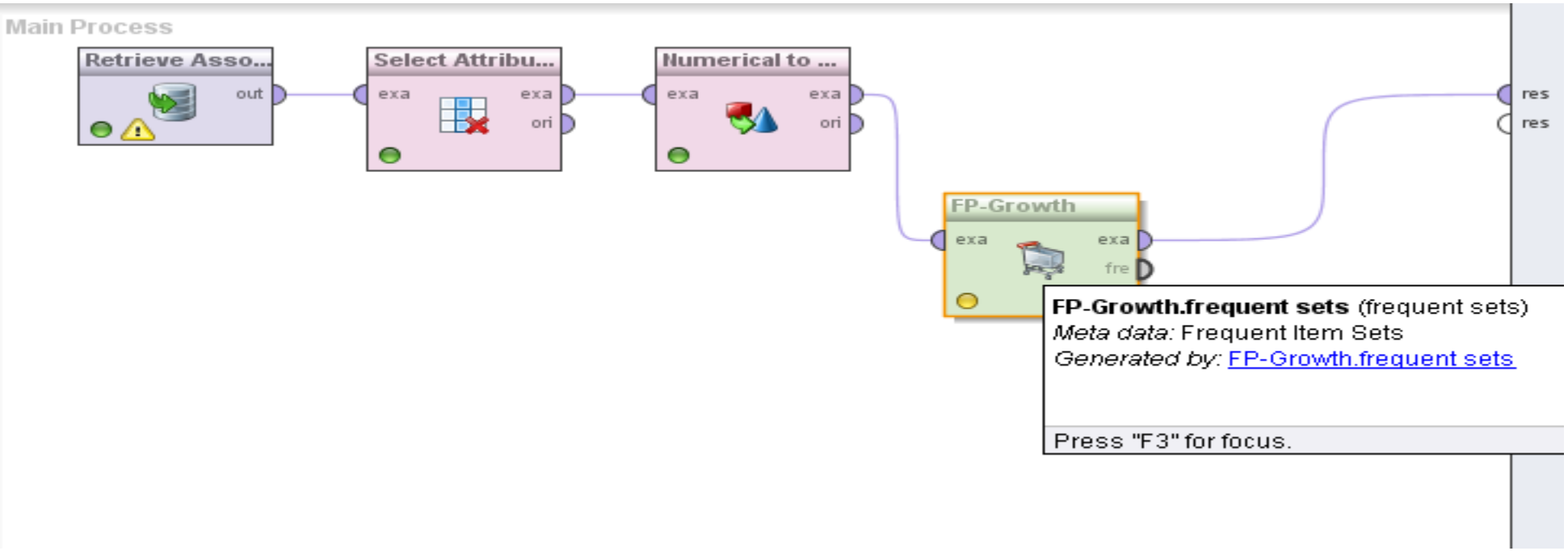
Role	Name	Type	Statistics	Range	Missings
regular	Family	binominal	mode = false (2125), least = true (1358)	false (2125), true (1358)	0.0
regular	Hobbies	binominal	mode = false (2438), least = true (1045)	false (2438), true (1045)	0.0
regular	Social_Club	binominal	mode = false (2828), least = true (655)	false (2828), true (655)	0.0
regular	Political	binominal	mode = false (3156), least = true (327)	false (3156), true (327)	0.0
regular	Professional	binominal	mode = false (2353), least = true (1130)	false (2353), true (1130)	0.0
regular	Religious	binominal	mode = false (2025), least = true (1458)	false (2025), true (1458)	0.0
regular	Support_Group	binominal	mode = false (2930), least = true (553)	false (2930), true (553)	0.0

(Design Perspective) Frequency Pattern Analysis: FP-Growth



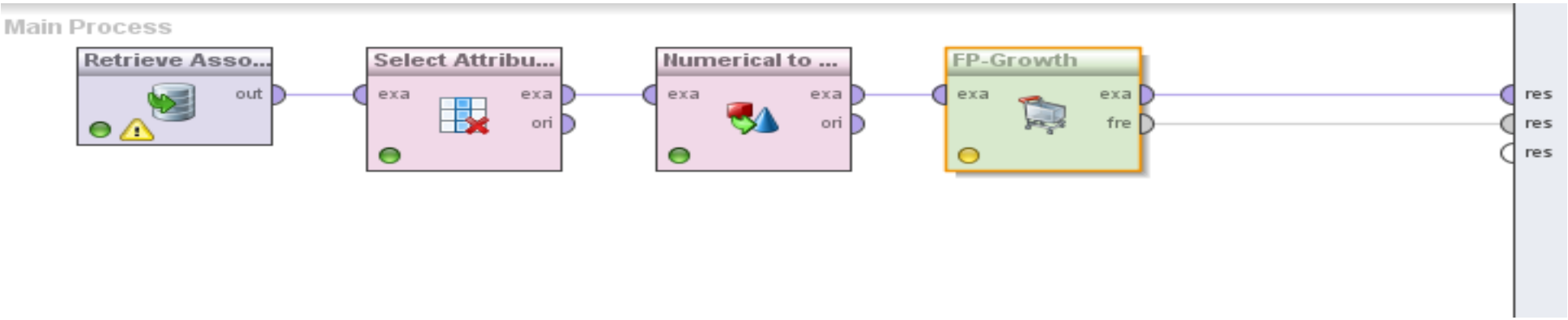


FP-Growth

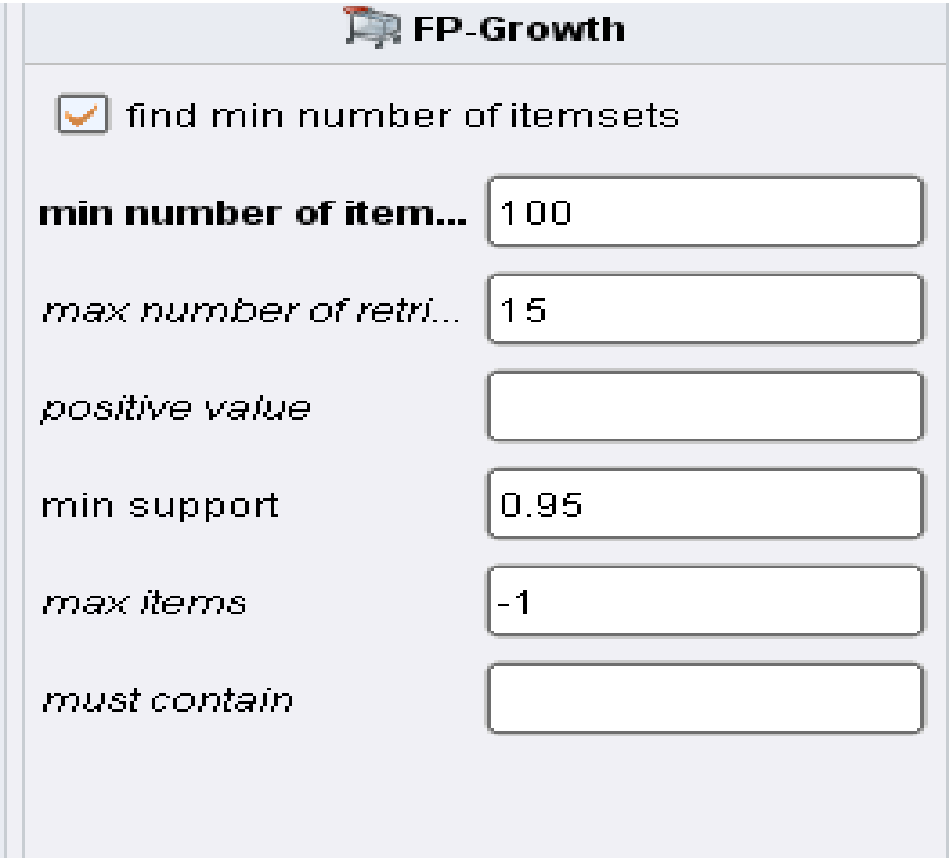

Main Process
Retrieve Asso...
out
Select Attribu...
exa
exa
ori
Numerical to ...
exa
exa
ori
FP-Growth
exa
exa
fre
res
res
FP-Growth.frequent sets (frequent sets)
Meta data: Frequent Item Sets
Generated by: FP-Growth.frequent sets
Press "F3" for focus.

FP-Growth - your DM stream

Both your *exa port and your fre port are connected to res ports*



Parameters pane



Play

Result Overview

Table View Annotations

No. of Sets: 6

Total Max. Size: 2

Min. Size: 1

Max. Size: 2

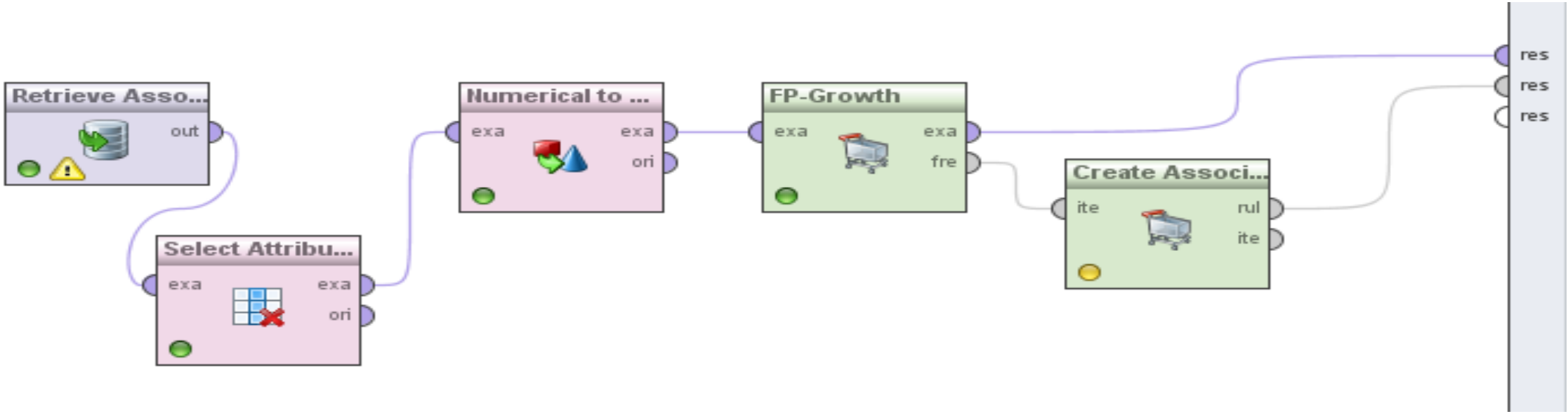
Contains Item:

Size	Support	Item 1	Item 2
1	0.419	Religious	
1	0.390	Family	
1	0.324	Professiona	
1	0.300	Hobbies	
2	0.225	Religious	Family
2	0.239	Religious	Hobbies

Religious organizations might have some natural connections with Family and Hobby organizations

Further investigation

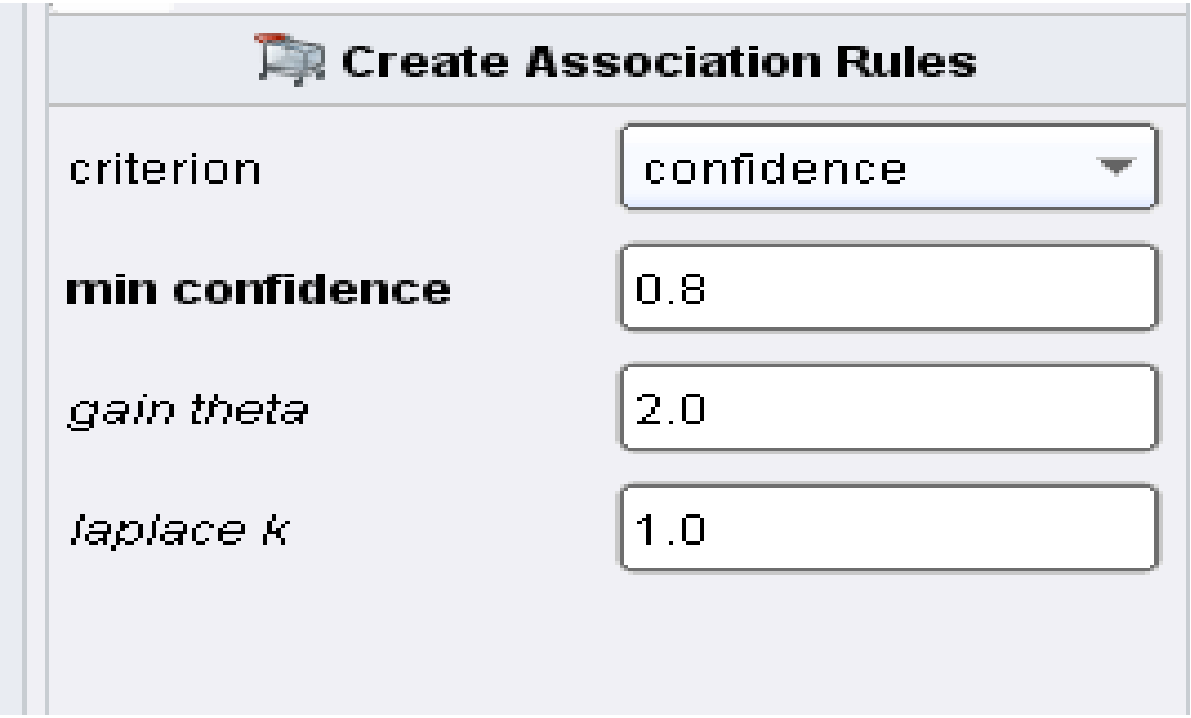
Create Association Rules operator


Retrieve Asso...
out
Select Attribu...
exa
exa
ori
Numerical to ...
exa
exa
ori
FP-Growth
exa
exa
fre
Create Associ...
ite
rul
ite
res
res
res

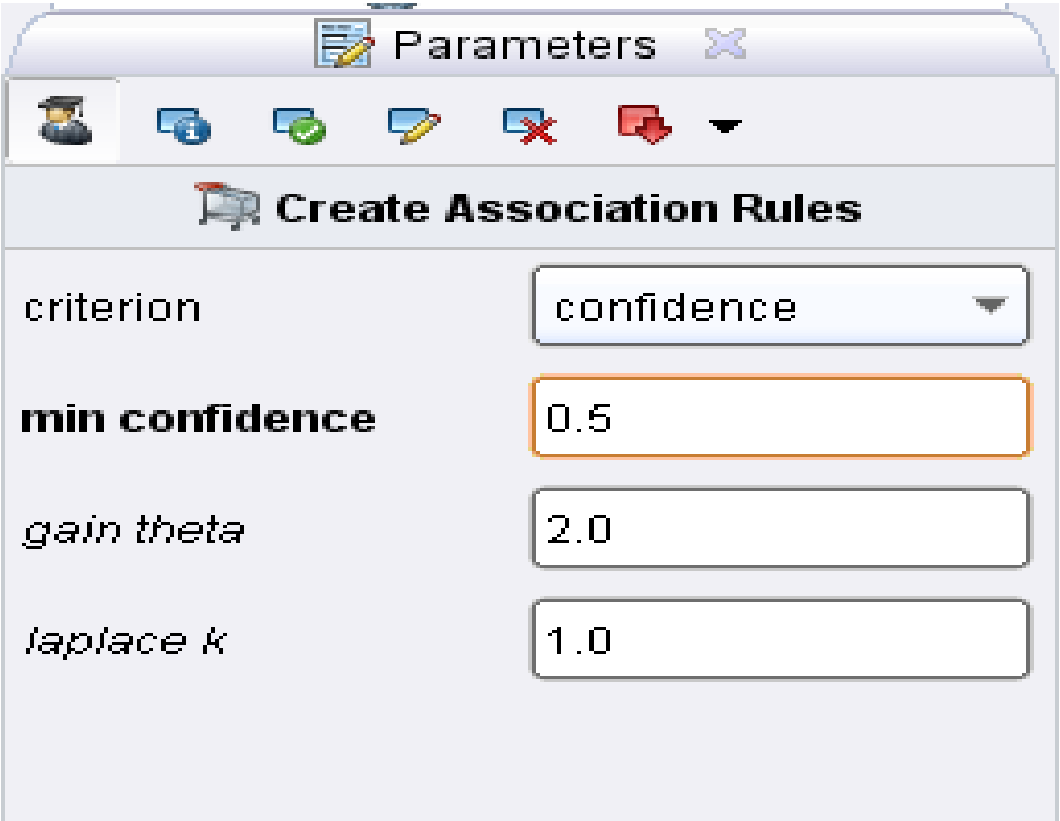
Play

Result Overview
AssociationRules (Create Association Rules)
Table View
Graph View
Text View
Annotations
no rules found

Parameters


Create Association Rules
criterion
confidence
min confidence
0.8
gain theta
2.0
laplace k
1.0

Changing the confidence percent threshold



Four rules found with the 50% confidence threshold

Result Overview | AssociationRules (Create Association Rules)

Table View | Graph View | Text View | Annotations

Show rules matching

all of these conclusions:

Religious
Family
Hobbies

No.	Premises	Conclusion	Support	Confid...	LaPl...	Gain	p-s	Lift	Conv...
1	Religious	Family	0.225	0.536	0.863	-0.61	0.061	1.376	1.316
2	Religious	Hobbies	0.239	0.571	0.873	-0.59	0.113	1.902	1.630
3	Family	Religious	0.225	0.576	0.881	-0.55	0.061	1.376	1.371
4	Hobbies	Religious	0.239	0.796	0.953	-0.36	0.113	1.902	2.852

No.	Premises	Conclusion	Support	Confid...	LaPl...	Gain	p-s	Lift	Conv...
1	Religious	Family	0.225	0.536	0.863	-0.61	0.061	1.376	1.316
2	Religious	Hobbies	0.239	0.571	0.873	-0.59	0.113	1.902	1.630
3	Family	Religious	0.225	0.576	0.881	-0.55	0.061	1.376	1.371
4	Hobbies	Religious	0.239	0.796	0.953	-0.36	0.113	1.902	2.852

Confidence - Support

1	2	3	4	5	6	7	8	9	10

Data set=10 baskets (transactions)

Cookies --> Milk (Cookies implies Milk)

$$\text{Confidence}=\frac{3\ (baskets\ \ with\ cookies\ \ and\ \ milk}{4\ (baskets\ \ with\ cookies\ )} = 0.75\ (75\%)$$

Milk --> Cookies

$$\text{Confidence}=\frac{3\ (baskets\ \ with\ cookies\ \ and\ \ milk\ )}{7\ (baskets\ \ with\ milk\ )} = 0.429\ (43\%)$$

Cookies --> Milk, Milk --> Cookies

$$\text{Support}=\frac{3\ (baskets\ \ with\ cookies\ \ and\ \ milk}{10\ (baskets\ )} = 0.3\ (30\%)$$

Confidence - Support

Γενίκευση

Confidence(X-->Y) is an estimation of the probability Pr(Y | X) of observing Y giving X.

Support(X-->Y) is the proportion of transactions in the data set which contain the item set (X,Y).

Support(X∪Y) is the number of transactions containing X and Y, support(X) is the number of transactions containing X, support(Y) is the number of transactions containing Y

Ισχύει:

Confidence(X-->Y)=$\frac{Support\ (X\ U\ Y)}{Support\ (X)}$

Lift - Conviction

Lift

The ratio of the observed support to that expected if X and Y were independent

$$\text{Lift}(X\text{-->}Y)=\frac{Support\ \ (X\ U\ Y)}{((Support\ \ (X)*(Support\ \ (Y))}$$

$$\text{Lift}(X\text{-->}Y)=\frac{Confident\ \ (X--->Y)}{Support\ \ (Y)}$$

Lift measures how far from independence are X and Y

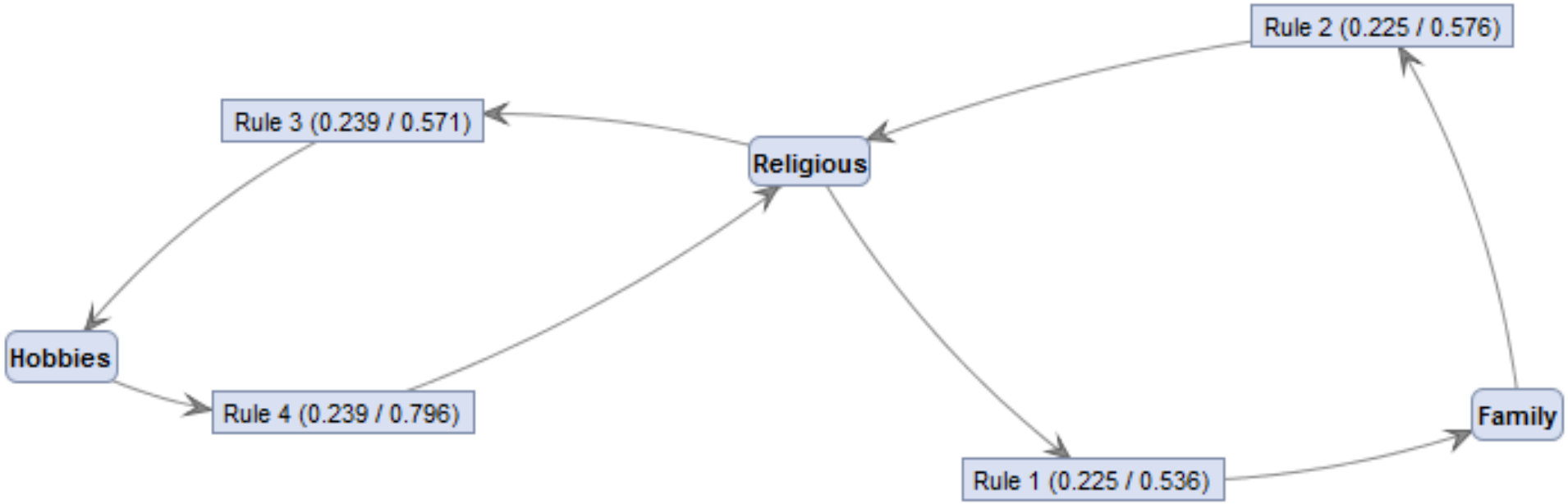
Lift ≈1 It implies that the rule is not interesting

Conviction

$$\text{Conviction}(X\text{-->}Y)=\frac{1-\text{support}\ \ (\text{Y})}{1-\text{confidence}\ \ \ (\text{X}--->\text{Y})}$$

Conviction is sensitive to rule direction i.e. conviction(X --> Y) <> conviction(Y --> X).

Plot view


Rule 2 (0.225 / 0.576)
Rule 3 (0.239 / 0.571)
Religious
Hobbies
Rule 4 (0.239 / 0.796)
Family
Rule 1 (0.225 / 0.536)

Text view

Association Rules

[Religious] --> [Family] (confidence: 0.536)
[Religious] --> [Hobbies] (confidence: 0.571)
[Family] --> [Religious] (confidence: 0.576)
[Hobbies] --> [Religious] (confidence: 0.796)

Υπάρχουν δεσμοί μεταξύ των τύπων κοινότητας; Ναι.

Διαπιστώσαμε ότι εκκλησίες, μέλη οικογενειακού τύπου οργανώσεων, και σύλλογοί με κοινό χόμπι της κοινότητας έχουν κάποια κοινά μέλη. Μπορεί να είναι λίγο παράξενο το γεγονός ότι οι πολιτικές και επαγγελματικές ομάδες δεν φαίνεται να αλληλοσυνδέονται, αλλά αυτές οι ομάδες είναι πιο εξειδικευμένες και επομένως μπορεί να μην έχουν μεγάλες ανάγκες για οργανωτική συνεργασία / συμπόρευση.

M. North, Data Mining for the Masses, 2012

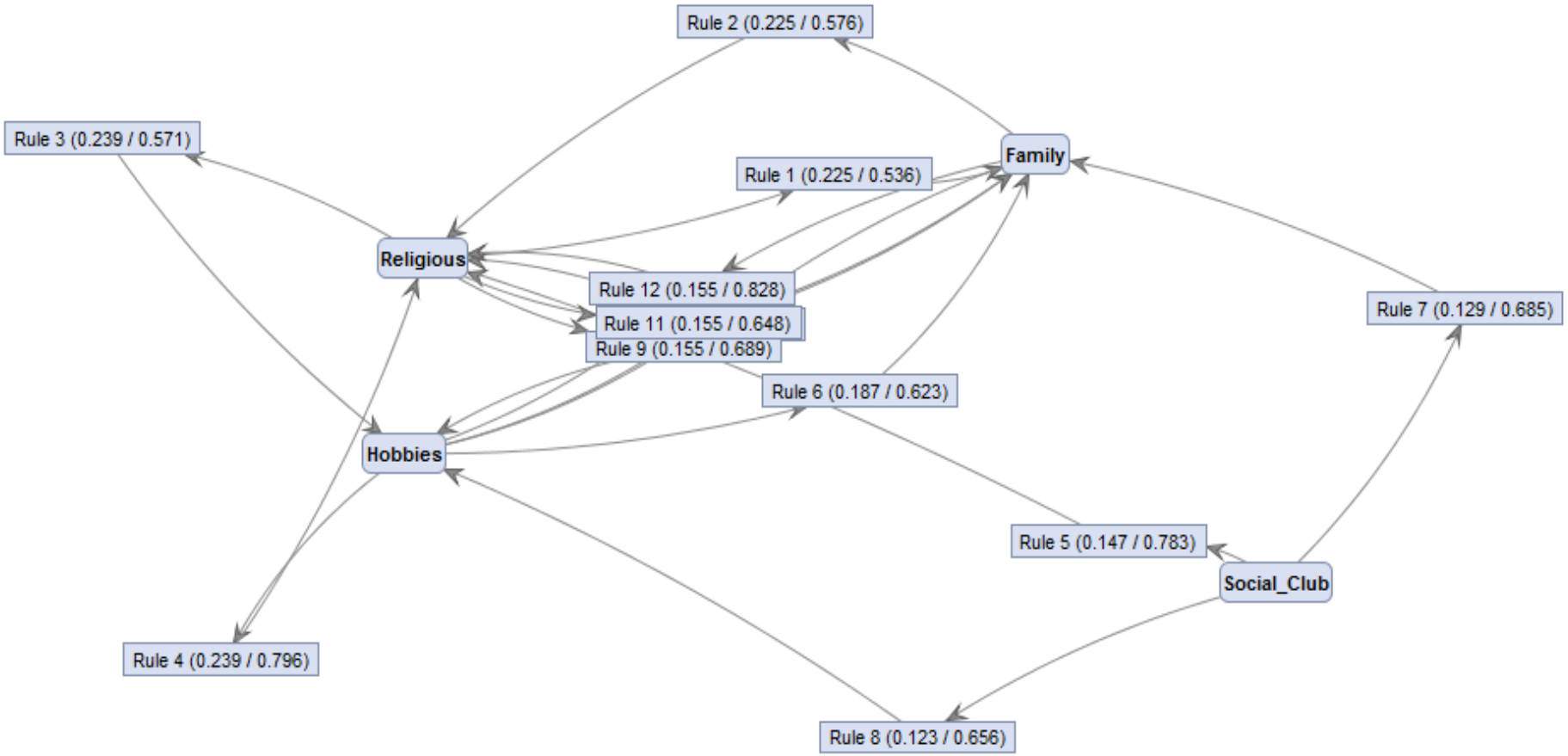
Lowering min support to .5

no	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	Hobbies	Religious, Family	0.155	0.516	0.888	-0.445	0.087	2.297	1.602
2	Religious	Family	0.225	0.536	0.863	-0.613	0.061	1.376	1.316
3	Religious	Hobbies	0.239	0.571	0.873	-0.598	0.113	1.902	1.630
4	Family	Religious	0.225	0.576	0.881	-0.555	0.061	1.376	1.371
5	Hobbies	Family	0.187	0.623	0.913	-0.413	0.070	1.598	1.618
6	Religious, Hobbies	Family	0.155	0.648	0.932	-0.323	0.062	1.662	1.732
7	Social_Club	Hobbies	0.123	0.656	0.946	-0.253	0.067	2.188	2.038
8	Social_Club	Family	0.129	0.685	0.950	-0.247	0.056	1.758	1.940
9	Religious, Family	Hobbies	0.155	0.689	0.943	-0.294	0.087	2.297	2.253
10	Social_Club	Religious	0.147	0.783	0.966	-0.229	0.069	1.871	2.682
11	Hobbies	Religious	0.239	0.796	0.953	-0.361	0.113	1.902	2.852
12	Family, Hobbies	Religious	0.155	0.828	0.973	-0.219	0.077	1.978	3.379

Size	Support	Item1	Item2	Item3
1	0.419	Religious		
1	0.390	Family		
1	0.324	Professional		
1	0.300	Hobbies		
1	0.188	Social_Club		
1	0.159	Support_Group		
2	0.225	Religious	Family	
2	0.124	Religious	Professional	
2	0.239	Religious	Hobbies	
2	0.147	Religious	Social_Club	
2	0.130	Family	Professional	
2	0.187	Family	Hobbies	
2	0.129	Family	Social_Club	
2	0.123	Hobbies	Social_Club	
3	0.155	Religious	Family	Hobbies

Association Rules

[Hobbies] --> [Religious, Family] (confidence: 0.516)
[Religious] --> [Family] (confidence: 0.536)
[Religious] --> [Hobbies] (confidence: 0.571)
[Family] --> [Religious] (confidence: 0.576)
[Hobbies] --> [Family] (confidence: 0.623)
[Religious, Hobbies] --> [Family] (confidence: 0.648)
[Social_Club] --> [Hobbies] (confidence: 0.656)
[Social_Club] --> [Family] (confidence: 0.685)
[Religious, Family] --> [Hobbies] (confidence: 0.689)
[Social_Club] --> [Religious] (confidence: 0.783)
[Hobbies] --> [Religious] (confidence: 0.796)
[Family, Hobbies] --> [Religious] (confidence: 0.828)


Rule 2 (0.225 / 0.576)
Rule 3 (0.239 / 0.571)
Family
Rule 1 (0.225 / 0.536)
Religious
Rule 12 (0.155 / 0.828)
Rule 11 (0.155 / 0.648)
Rule 9 (0.155 / 0.689)
Rule 7 (0.129 / 0.685)
Rule 6 (0.187 / 0.623)
Hobbies
Rule 5 (0.147 / 0.783)
Social_Club
Rule 4 (0.239 / 0.796)
Rule 8 (0.123 / 0.656)

ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ