Subject

In many situations, we want to characterize a subset of the dataset. The GROUP CHARACTERIZATION component allows comparing several subgroups, it computes and compares descriptive statistics on the subsets. But, this component performs univariate analysis. It uses individually the attributes and does not analyze the possible interaction between two or more variables.

In this tutorial, we show a new component SPV ASSOC TREE that allows characterizing a subset of examples with the conjunction of variables. In fact, it is a "supervised like" association rule algorithm where we define the consequent of the rule.

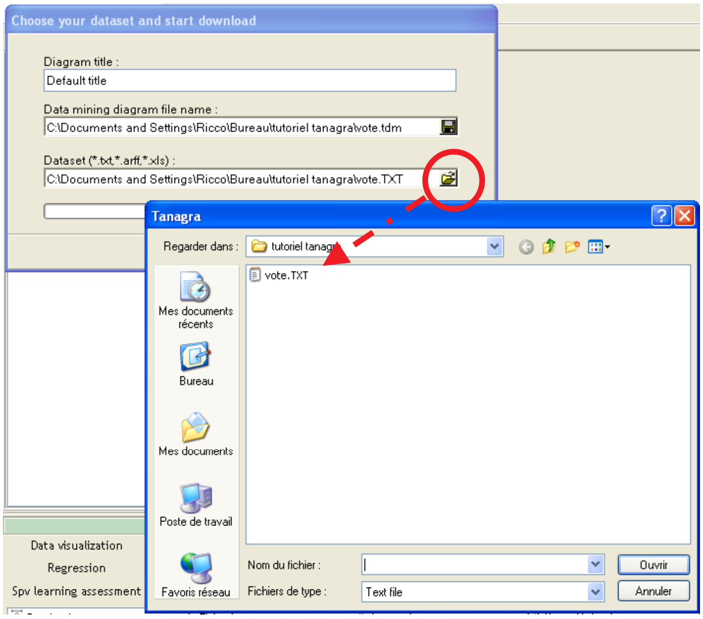
Dataset

We use the "Congress vote" dataset (VOTE.TXT). We want to characterize the republicans' subgroup from their votes on several subjects.

Multivariate group characterization

Importation

First of all, we create a new diagram and import the dataset with the FILE / NEW menu.



Choose your dataset and start download

Diagram title : Default title

Data mining diagram file name : C:\Documents and Settings\Ricco\Bureau\tutoriel tanagra\vote.tdm

Dataset (*.txt,*.arff,*.xls) : C:\Documents and Settings\Ricco\Bureau\tutoriel tanagra\vote.TXT

Tanagra

Regarder dans : tutoriel tanagra

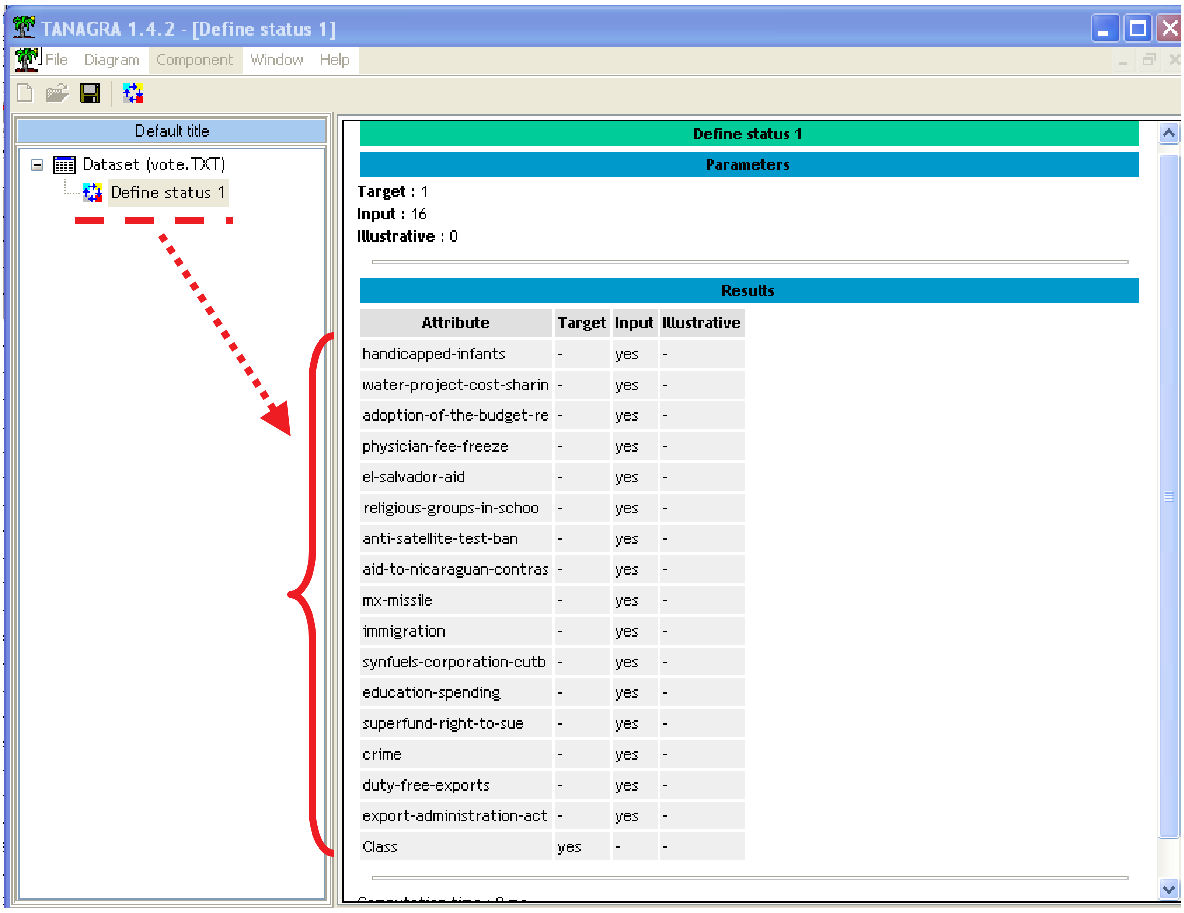
vote.TXT

Nom du fichier :

Fichiers de type : Text file

Select attributes

We add the DEFINE STATUS component; we set as TARGET the CLASS attribute, and set as INPUT the others.


TANAGRA 1.4.2 - [Define status 1]
File Diagram Component Window Help
Default title
Dataset (vote.TXT)
Define status 1

Define status 1

Parameters

Target : 1
Input : 16
Illustrative : 0

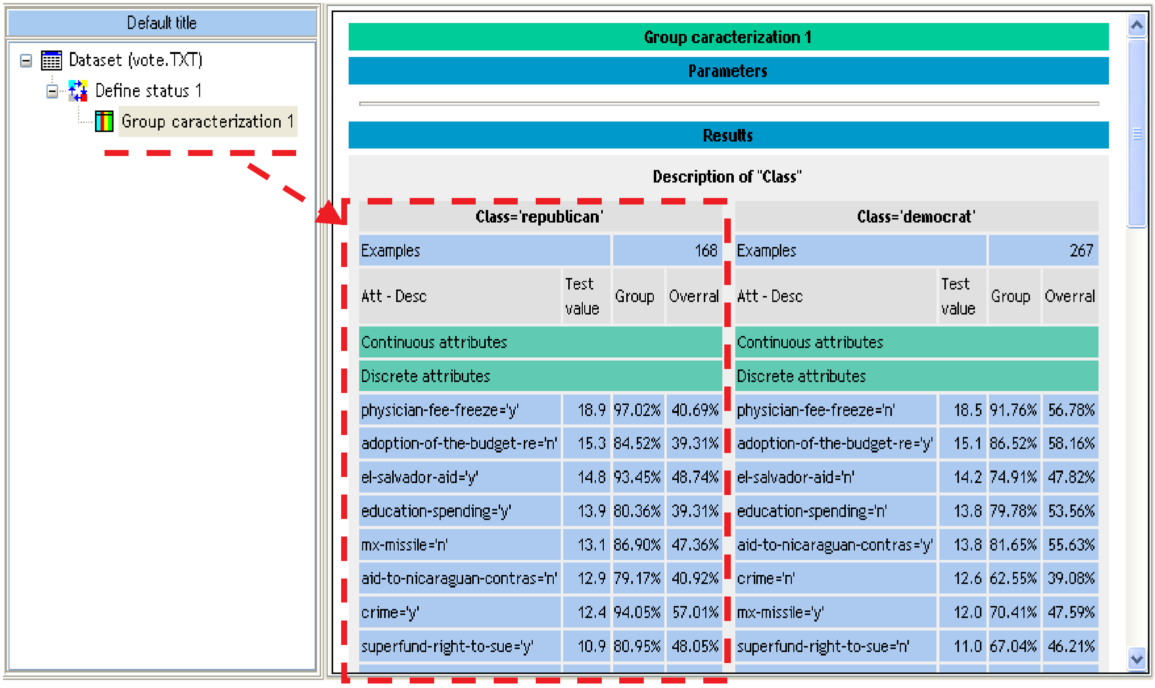
Results

Attribute	Target	Input	Illustrative
handicapped-infants	-	yes	-
water-project-cost-sharin	-	yes	-
adoption-of-the-budget-re	-	yes	-
physician-fee-freeze	-	yes	-
el-salvador-aid	-	yes	-
religious-groups-in-schoo	-	yes	-
anti-satellite-test-ban	-	yes	-
aid-to-nicaraguan-contras	-	yes	-
mx-missile	-	yes	-
immigration	-	yes	-
synfuels-corporation-cutb	-	yes	-
education-spending	-	yes	-
superfund-right-to-sue	-	yes	-
crime	-	yes	-
duty-free-exports	-	yes	-
export-administration-act	-	yes	-
Class	yes	-	-

Univariate group characterization

We use the GROUP CHARACTERIZATION component to compare the conditional probabilities in each subgroup. The attributes are ranked according their test-value that indicates the importance of the difference between groups.

Our goal is to study the republicans' subgroup. But, this component shows the results on each subgroup, including the democrats. We read only the first column.



Group caracterization 1

Parameters

Results

Description of "Class"

Class='republican'				Class='democrat'			
Examples			168	Examples			267
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes				Continuous attributes			
Discrete attributes				Discrete attributes			
physician-fee-freeze='y'	18.9	97.02%	40.69%	physician-fee-freeze='n'	18.5	91.76%	56.78%
adoption-of-the-budget-re='n'	15.3	84.52%	39.31%	adoption-of-the-budget-re='y'	15.1	86.52%	58.16%
el-salvador-aid='y'	14.8	93.45%	48.74%	el-salvador-aid='n'	14.2	74.91%	47.82%
education-spending='y'	13.9	80.36%	39.31%	education-spending='n'	13.8	79.78%	53.56%
mx-missile='n'	13.1	86.90%	47.36%	aid-to-nicaraguan-contras='y'	13.8	81.65%	55.63%
aid-to-nicaraguan-contras='n'	12.9	79.17%	40.92%	crime='n'	12.6	62.55%	39.08%
crime='y'	12.4	94.05%	57.01%	mx-missile='y'	12.0	70.41%	47.59%
superfund-right-to-sue='y'	10.9	80.95%	48.05%	superfund-right-to-sue='n'	11.0	67.04%	46.21%

There are 435 congressmen, 168 are republicans (38%). The first attribute that characterizes the republicans is PHYSICIAN-FEE-FREEZE: 40.69% (177 congressmen) have answer YES; among the republicans, this proportion is increased to 97.02%, i.e. 97.02% x 168 = 163 congressmen.

We can deduce several probabilities from theses results:

$$P(physician = y / republican) = 97.02\%$$

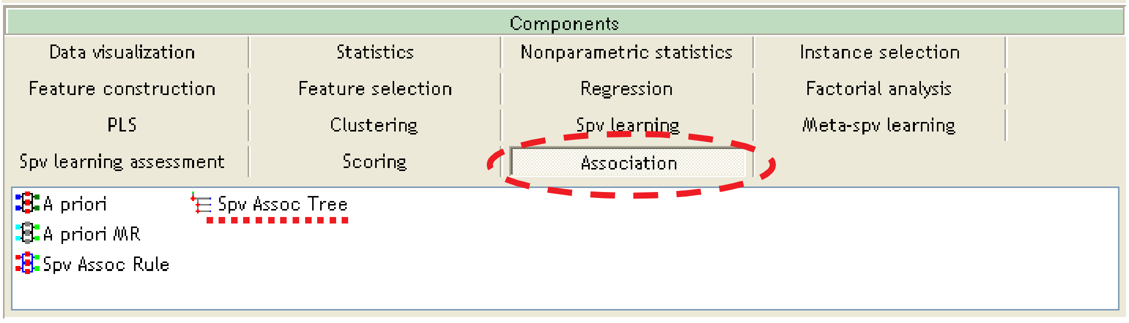
and

$$P(republican / physician = y) = \frac{P(republican \cap physician = y)}{P(physician = y)} = \frac{163}{177} = 92.1\%$$

We can read the same results on each attribute.

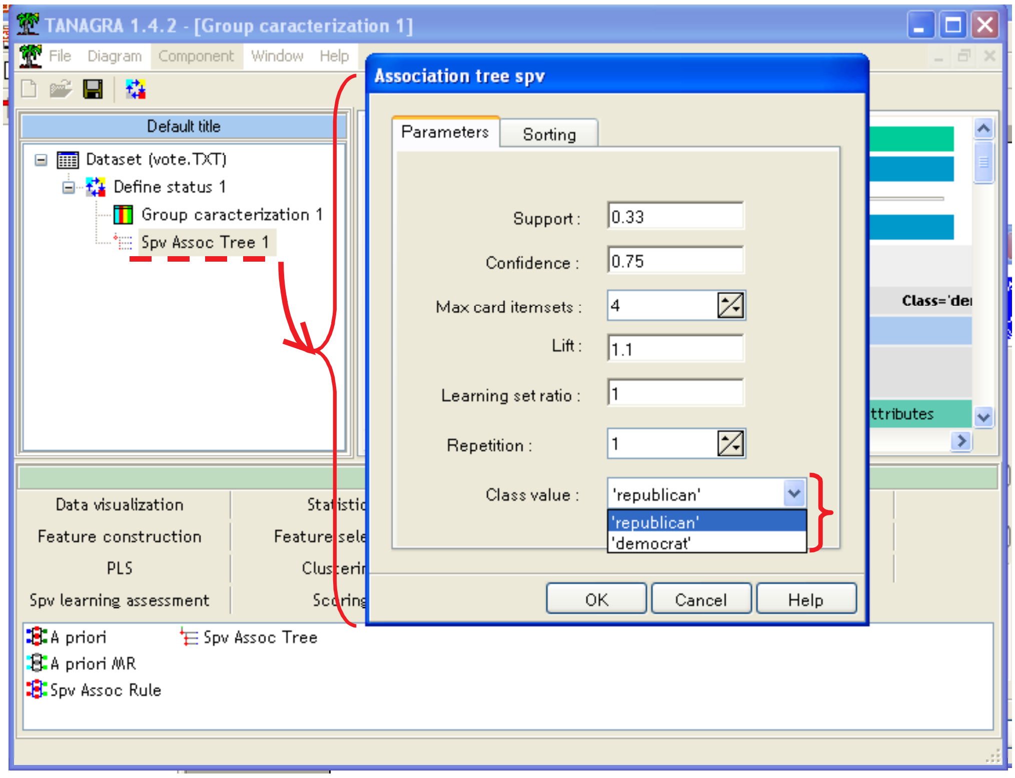
Multivariate characterization

This first component cannot handle the simultaneous effect of two or more variables. The characterization is not powerful.



The SPV ASSOC TREE component build association rule where we can define the item that we want obtains in the consequent of the rule. We use this one to characterize the republican's subgroup.

We add this component in the diagram. Let we see its main parameters.



Support : 0.33
Confidence : 0.75
Max card itemsets : 4
Lift : 1.1
Learning set ratio : 1
Repetition : 1
Class value : 'republican'
'republican'
'democrat'

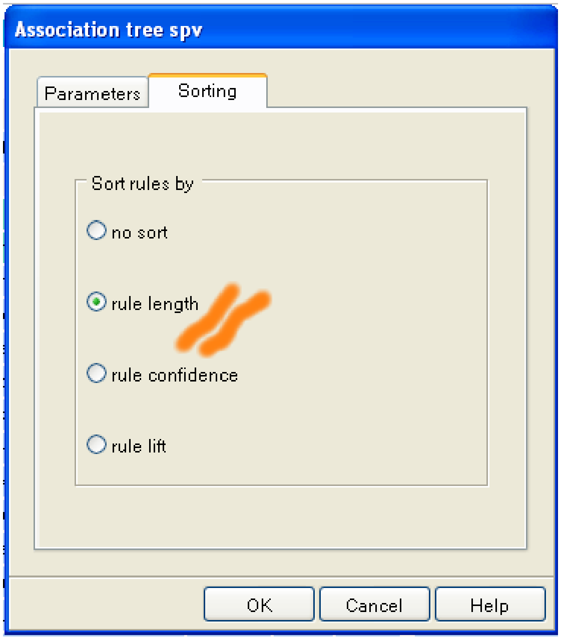
The first four parameters (SUPPORT, CONFIDENCE, MAX ITEMSET, LIFT) are the same one that the standard association rules A PRIORI algorithm. They enable to restrict the number of produced rules.

LEARNING SET RATIO enables to create learning and test set from the whole dataset. It is not useful for our problem.

REPETITION is an experimental parameter, which we can ignore.

The last parameter is very important, CLASS VALUE enables us to define the subgroup which we want characterize. We set the "republican" value here.

In the SORTING tab, we set the criterion to use when we sort the rules. It can be of primary importance if we obtain a lot of rules.



Sort rules by

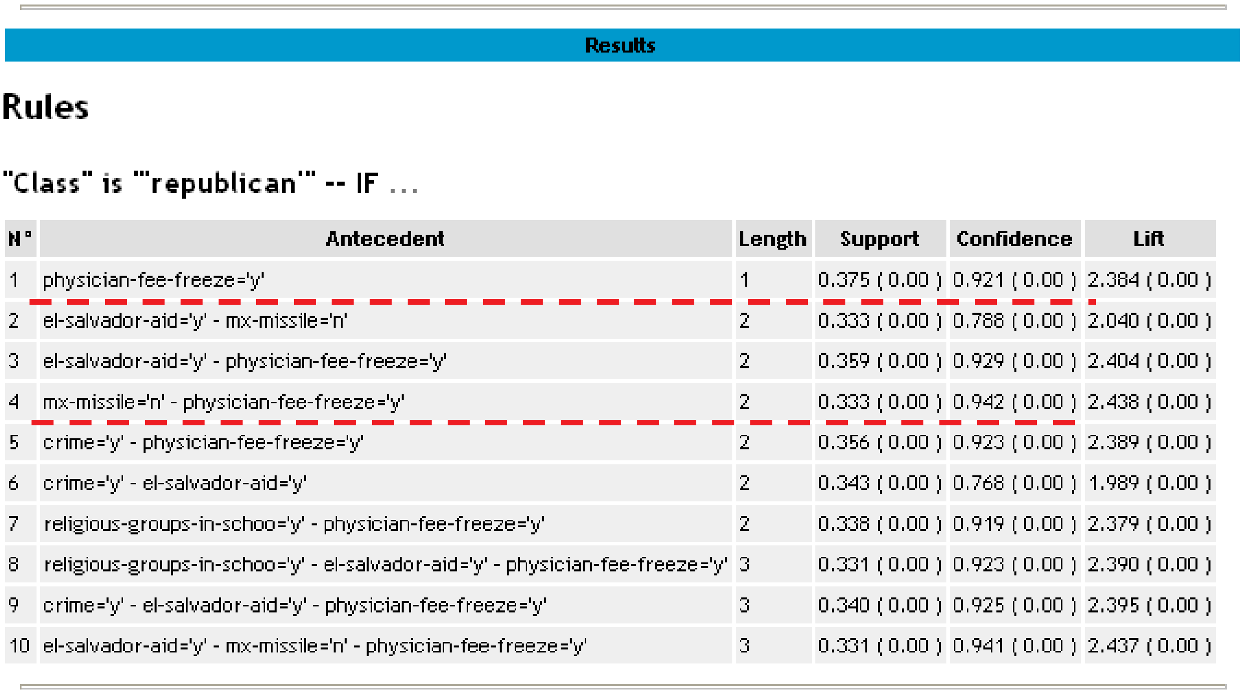
○ no sort

● rule length

○ rule confidence

○ rule lift

We obtain the following results.



Results

Rules

"Class" is "'republican'" -- IF ...

N°	Antecedent	Length	Support	Confidence	Lift
1	physician-fee-freeze='y'	1	0.375 (0.00)	0.921 (0.00)	2.384 (0.00)
2	el-salvador-aid='y' - mx-missile='n'	2	0.333 (0.00)	0.788 (0.00)	2.040 (0.00)
3	el-salvador-aid='y' - physician-fee-freeze='y'	2	0.359 (0.00)	0.929 (0.00)	2.404 (0.00)
4	mx-missile='n' - physician-fee-freeze='y'	2	0.333 (0.00)	0.942 (0.00)	2.438 (0.00)
5	crime='y' - physician-fee-freeze='y'	2	0.356 (0.00)	0.923 (0.00)	2.389 (0.00)
6	crime='y' - el-salvador-aid='y'	2	0.343 (0.00)	0.768 (0.00)	1.989 (0.00)
7	religious-groups-in-schoo='y' - physician-fee-freeze='y'	2	0.338 (0.00)	0.919 (0.00)	2.379 (0.00)
8	religious-groups-in-schoo='y' - el-salvador-aid='y' - physician-fee-freeze='y'	3	0.331 (0.00)	0.923 (0.00)	2.390 (0.00)
9	crime='y' - el-salvador-aid='y' - physician-fee-freeze='y'	3	0.340 (0.00)	0.925 (0.00)	2.395 (0.00)
10	el-salvador-aid='y' - mx-missile='n' - physician-fee-freeze='y'	3	0.331 (0.00)	0.941 (0.00)	2.437 (0.00)

We find again similar results to the GROUP CHARACTERIZATION component, but we have also more detailed results.

The rule n°1 is the same one, with another kind of measures:

- The SUPPORT shows $P(physician = y \cap republican) = \frac{163}{435} = 37.5\%$
- The CONFIDENCE shows the probability $P(republican / physician = y) = 92.1\%$

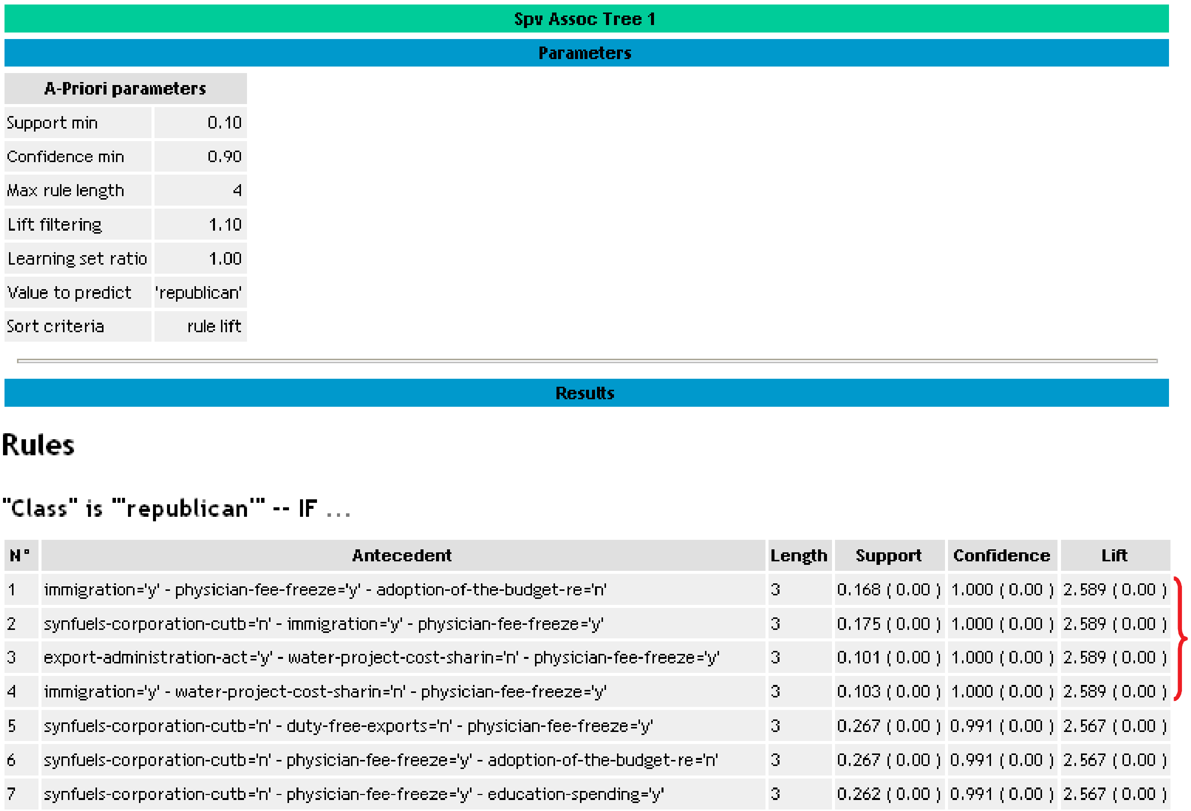
- The LIFT computes $\frac{P(republican / physician = y)}{P(republican)} = \frac{92.1\%}{38.6\%} = 2.384$

The main contribution of this new component is that we can handle the interaction between two or more attributes. We see for instance the rule n°4. We see the contribution of MX-MISSILE to PHYSICIAN in the characterization. We obtain a more accurate rule, the confidence of the rule becomes: $P(republican / physician = y \cap mx - missile = n) = 94.2\%$. Nevertheless, this improvement seems to be not significant. We see in the next section if we can obtain more convincing results.

Modify the parameters

This is a powerful approach, but the results rely heavily on the parameters. If we are too restrictive, we obtain a few number of rules, and we can miss some interesting rules; if we are too permissive, we obtain a great number of rules, it becomes very hard to distinguish the interesting ones.

We try to reduce the SUPPORT MIN to 10%, and increase the CONFIDENCE MIN to 90%; we sort the rules according to the LIFT.



Spv Assoc Tree 1

Parameters

A-Priori parameters	
Support min	0.10
Confidence min	0.90
Max rule length	4
Lift filtering	1.10
Learning set ratio	1.00
Value to predict	'republican'
Sort criteria	rule lift

Results

Rules

"Class" is "'republican'" -- IF ...

N°	Antecedent	Length	Support	Confidence	Lift
1	immigration='y' - physician-fee-freeze='y' - adoption-of-the-budget-re='n'	3	0.168 (0.00)	1.000 (0.00)	2.589 (0.00)
2	synfuels-corporation-cutb='n' - immigration='y' - physician-fee-freeze='y'	3	0.175 (0.00)	1.000 (0.00)	2.589 (0.00)
3	export-administration-act='y' - water-project-cost-sharin='n' - physician-fee-freeze='y'	3	0.101 (0.00)	1.000 (0.00)	2.589 (0.00)
4	immigration='y' - water-project-cost-sharin='n' - physician-fee-freeze='y'	3	0.103 (0.00)	1.000 (0.00)	2.589 (0.00)
5	synfuels-corporation-cutb='n' - duty-free-exports='n' - physician-fee-freeze='y'	3	0.267 (0.00)	0.991 (0.00)	2.567 (0.00)
6	synfuels-corporation-cutb='n' - physician-fee-freeze='y' - adoption-of-the-budget-re='n'	3	0.267 (0.00)	0.991 (0.00)	2.567 (0.00)
7	synfuels-corporation-cutb='n' - physician-fee-freeze='y' - education-spending='y'	3	0.262 (0.00)	0.991 (0.00)	2.567 (0.00)

There are 270 rules! The first four ones are very accurate; they have not a counter-examples.