## 1 Subject

# Association Rules mining with Tanagra, R (arules package), Orange, RapidMiner, Knime and Weka.

This document extends a previous tutorial dedicated to the comparison of various implementations of association rules mining<sup>1</sup>. We had analyzed Tanagra, Orange and Weka. We extend here the comparison to R, RapidMiner and Knime.

We handle an attribute-value dataset. It is not the usual data format for the association rule mining where the "native" format is rather the transactional database. We see in this tutorial than some of tools can automatically recode the data. Others require an explicit transformation. Thus, we must find the right components and the correct sequence of treatments to produce the transactional data format. The process is not always easy according to the software.

#### 2 Dataset and tools

We use the CREDIT-GERMAN.TXT<sup>2</sup> dataset. The characteristics of 1000 customers of a finance company are described. It comes from the UCI server<sup>3</sup>. The continuous attributes are discretized. We show below a sample of the dataset.

| 2    | credit-german.xk | 5             |                         |                     |             |                  |            |                    | <  |
|------|------------------|---------------|-------------------------|---------------------|-------------|------------------|------------|--------------------|----|
|      | A                | В             | С                       | D                   | E           | F                | G          | н                  | -  |
| 1    | checking_status  | disc_duration | credit_history          | purpose             | disc_amount | savings_status   | employment | personal_status    | -  |
| 2    | <0               | lo_1_year     | critical/other existing | radio/tv            | 1000_2000   | no known savings | >=7        | male single        |    |
| 3    | 0<=X<200         | up_2_years    | existing paid           | radio/tv            | up_2000     | <100             | 1<=X<4     | female div/dep/mar |    |
| 4    | no checking      | lo_1_year     | critical/other existing | education           | up_2000     | <100             | 4<=X<7     | male single        |    |
| 5    | <0               | up_2_years    | existing paid           | furniture/equipment | up_2000     | <100             | 4<=X<7     | male single        |    |
| 6    | <0               | 1_2_years     | delayed previously      | new car             | up_2000     | <100             | 1<=X<4     | male single        |    |
| 7    | no checking      | up_2_years    | existing paid           | education           | up_2000     | no known savings | 1<=X<4     | male single        |    |
| 8    | no checking      | 1_2_years     | existing paid           | furniture/equipment | up_2000     | 500<=X<1000      | >=7        | male single        |    |
| 9    | 0<=X<200         | up_2_years    | existing paid           | used car            | up_2000     | <100             | 1<=X<4     | male single        |    |
| 10   | no checking      | lo_1_year     | existing paid           | radio/tv            | up_2000     | >=1000           | 4<=X<7     | male div/sep       |    |
| 11   | 0<=X<200         | up_2_years    | critical/other existing | new car             | up_2000     | <100             | unemployed | male mar/wid       |    |
| 12   | 0<=X<200         | lo_1_year     | existing paid           | new car             | 1000_2000   | <100             | <1         | female div/dep/mar |    |
| 13   | <0               | up_2_years    | existing paid           | business            | up_2000     | <100             | <1         | female div/dep/mar |    |
| 14   | 0<=X<200         | lo_1_year     | existing paid           | radio/tv            | 1000_2000   | <100             | 1<=X<4     | female div/dep/mar |    |
| 15   | <0               | 1_2_years     | critical/other existing | new car             | 1000_2000   | <100             | >=7        | male single        |    |
| 16   | <0               | 1_2_years     | existing paid           | new car             | 1000_2000   | <100             | 1<=X<4     | female div/dep/mar |    |
| 17   | <0               | 1_2_years     | existing paid           | radio/tv            | 1000_2000   | 100<=X<500       | 1<=X<4     | female div/dep/mar |    |
| 18   | no checking      | 1_2_years     | critical/other existing | radio/tv            | up_2000     | no known savings | >=7        | male single        |    |
| 19   | <0               | up_2_years    | no credits/all paid     | business            | up_2000     | no known savings | <1         | male single        |    |
| 20   | 0<=X<200         | 1_2_years     | existing paid           | used car            | up_2000     | <100             | >=7        | female div/dep/mar |    |
| 21   | no checking      | 1_2_years     | existing paid           | radio/tv            | up_2000     | 500<=X<1000      | >=7        | male single        |    |
| 22   | no checking      | lo_1_year     | critical/other existing | new car             | up_2000     | <100             | 1<=X<4     | male single        |    |
| 23   | <0               | lo_1_year     | existing paid           | radio/tv            | up_2000     | 500<=X<1000      | 1<=X<4     | male single        | -  |
| 14 - | 🕞 🕨 🔿 credit-ge  | erman /       |                         |                     | •           |                  |            |                    | 11 |

All the tools analyzed in this tutorial can handle parameters on the confidence (0.75) and support (0.25). We can set also the maximum number of items in a rule (10). The tools studied in this tutorial

<sup>&</sup>lt;sup>1</sup> <u>http://data-mining-tutorials.blogspot.com/2008/10/association-rule-learning.html</u>

<sup>&</sup>lt;sup>2</sup> <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit-german.zip</u>

<sup>&</sup>lt;sup>3</sup> <u>http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)</u>

are: Tanagra 1.4.28, R 2.7.2 (arules package 0.6-6), Orange 1.0b2, RapidMiner Community Edition, Knime 1.3.5 and Weka 3.5.6. These programs load the data and perform the calculations in memory. When the size of the database increases, the real bottleneck is the memory available on our personal computer.

### 3 Tanagra (A Priori component)

**Data importation and diagram initialization**. After launching Tanagra, we activate the FILE / NEW menu. We select the CREDIT-GERMAN.TXT data file.



The dataset contains 19 variables and 1000 instances.

**Specifying the status of the variables**. In order to specify the status of each descriptor in the analysis, we use the DEFINE STATUS component. We set all the variables as INPUT.

| 🕱 TANAGRA 1.4.28 - [Dataset (credit-german  | .txt)]   |             |
|---|--|-------------|
| 📅 File Diagram Component Window Help  |  | - 8 ×       |
|   |  |             |
| Default title   |  |             |
| Dataset (credit-german.txt)   | Define attribute statuses  | -german.cxt |
| Data visualization       Statistics         Feature selection       Regression         Spv learning       Meta-spv learning         Correlation scatterplot       Scatterplot with         Export dataset       Scatterplot         Scatterplot       Scatterplot | Parameters         Attributes :         Image:         Image: |             |
|   |  |             |

**Extraction of association rules**. We can insert the A PRIORI component now (ASSOCIATION tab). We activate the PARAMETERS menu in order to specify the parameters of the analysis.

| 💯 TANAGRA 1.4.28 - [Dataset (cred   | lit-german.txt)]     |   |                                  |
|---|----------------------|---|----------------------------------|
| Tile Diagram Component Window H   | Help                 |   | _ @ ×                            |
| 🗅 📽 📕   🎎   |                      |   |                                  |
| Default title   |                      | D:\DataMining\Databases_for_mining\comparison_TOW\asso                      | ciation_rule\credit-german.txt 🔼 |
| 🖃 🏢 Dataset (credit-german.txt)   |                      |   |                                  |
| 🖃 🎎 Define status 1   | (                    | Association rule parameter  |                                  |
| A priori 1<br>Parameters.<br>Execute<br>View  |                      | Parameters<br>Support : 0.25<br>Confidence : 0.75<br>Max card itemsets : 10 |                                  |
| Data visualization St:  | atistics Nonpa       |   | construction                     |
| Feature selection Res   | gression Fac         |   | lustering                        |
| Spv learning 📕 Meta-s   | spv learning Spv lea | OK Cancel Help  | sociation                        |
| A priori MR     Spv Assoc Outlier     A priori MR     Spv Assoc Rule     A priori PT     E Spv Assoc Tree |                      |   |                                  |

The LIFT criterion is set to 0, thus it does not influence the extraction process. Our results will be directly comparable to the results of other tools which do not handle this parameter.

We click on VIEW menu.

| 💯 TANAGRA 1.4.28 - [A   | priori 1]                              |          |                  |              |                    |                      |     |
|---|--|----------|------------------|--------------|--------------------|----------------------|-----|
| 💇 File Diagram Componen   | t Window Help                          |          |                  |              |                    | -                    | a × |
| D 📽 🖬   👪   |  |          |                  |              |                    |                      |     |
| Default title   | )                                      | Transa   | ctions 1000      | -,           |                    |                      | ^   |
| 🖃 🧱 Dataset (credit-ger   | man.txt)                               | Co       | unting items     | 1 - C        |                    |                      |     |
| Define status 1   |  | All item | s 71             |              |                    |                      |     |
|   |  | Filtered | ditems 31        | 1            |                    |                      |     |
|   |  | Cour     | nting itemsets   |              |                    |                      |     |
|   |  | card(it  | emset) = 2 162   |              |                    |                      |     |
|   |  | card(it  | emset) = 3 345   | 1            |                    |                      |     |
|   |  | card(it  | emset) = 4 334   | - i          |                    |                      |     |
|   |  | card(it  | emset) = 5 140   | 1            |                    |                      |     |
|   |  | card(it  | emset) = 6 16    |              |                    |                      |     |
|   |  |          | Rules            | i i          |                    |                      |     |
|   |  | Numbe    | er of rules 2986 | /            |                    |                      | ~   |
|   |  |          | Con              | nponents     |                    |                      |     |
| Data visualization  | Statistics                             |          | Nonparametri     | c statistics | Instance selection | Feature construction |     |
| Feature selection   | Regression                             |          | Factorial        | analysis     | PLS                | Clustering           |     |
| Spv learning  | Meta-spv learr                         | ning     | Spv learning a   | assessment   | Scoring            | Association          |     |
| 8 A priori : Assi<br>8 A priori MR 8 Spv<br>8 A priori PT ≒ Spv | oc Outlier<br>Assoc Rule<br>Assoc Tree |          |                  |              |                    |                      |     |
|   |  |          |                  |              |                    |                      |     |

Various indications are available: there are 71 items (attribute-value pair) into the dataset; 31 of them have a support  $\geq 0.25$ ; we see the number of itemsets of same length i.e. 162 itemsets with length = 2, etc.; thus, 2986 rules are extracted.

The rules are enumerated in the low part of the report. They are ranked according a decreasing value of the LIFT criterion.

|    | Number of rules : 2986   |  |       |         |            |  |  |  |  |  |  |
|----|--|--|-------|---------|------------|--|--|--|--|--|--|
| N" | Antecedent   | Consequent   | Lift  | Support | Confidence |  |  |  |  |  |  |
| 1  | "other_payment_plans=none" -<br>"existing_credits=one" -<br>"own_telephone=none"   | "credit_history=existing paid"   | 1.551 | 0.263   | 0.822      |  |  |  |  |  |  |
| 2  | "num_dependents=one" - "class=good" -<br>"credit_history=existing paid"            | "other_payment_plans=none" -<br>"existing_credits=one"                           | 1.548 | 0.253   | 0.808      |  |  |  |  |  |  |
| 3  | "other_parties=none" -<br>"num_dependents=one" -<br>"credit_history=existing paid" | "foreign_worker=yes" -<br>"other_payment_plans=none" -<br>"existing_credits=one" | 1.534 | 0.319   | 0.769      |  |  |  |  |  |  |
| 4  | "class=good" - "credit_history=existing paid"                                      | "other_payment_plans=none" -<br>"existing_credits=one"                           | 1.534 | 0.289   | 0.801      |  |  |  |  |  |  |
| 5  | "own_telephone=none" -<br>"credit_history=existing paid"                           | "other_payment_plans=none" -<br>"existing_credits=one"                           | 1.527 | 0.263   | 0.797      |  |  |  |  |  |  |
|    |  |  |       |         |            |  |  |  |  |  |  |

### 4 Tanagra (A Priori PT component)

There is a second component dedicated to the association rule extraction in Tanagra. This is an external program "apriori.exe" of Christian BORGELT (<u>http://www.borgelt.net/apriori.html</u>). We have already written a first tutorial where we show the functioning of this tool previously (<u>http://data-mining-tutorials.blogspot.com/2008/11/association-rule-learning-using-prefix.html</u>). BORGELT's program is really fast. But there is a slight limitation: only the rules with one item into the consequent can be generated. Therefore, for the same parameters above we obtain fewer rules.

In the previous diagram into Tanagra, we insert the A PRIORI PT component (ASSOCIATION tab). We set the parameters as follows.



We can click now on the VIEW menu.

| 🎬 A priori PT 1   |                              |         |       |        |   |  |  |  |
|---|------------------------------|---------|-------|--------|---|--|--|--|
| Execution log   |                              |         |       |        |   |  |  |  |
| D:\Temp\Exe\exe\apriori.exe - find association rules with the apriori algorithm<br>version 4.31 (2007.03.12) (c) 1996-2007 Christian Borgelt<br>reading C:\DOCUME~1\Maison\LOCALS~1\Temp\dat8D.tmp [71 item(s), 1000 transaction(s)] done [0.08s].<br>filtering, sorting and recoding items [31 item(s)] done [0.00s].<br>creating transaction tree done [0.00s].<br>checking subsets of size 1 2 3 4 5 6 done [0.01s].<br>writing D:\DataMining\Databases_for_mining\comparison_TOW\association_rule\output.rul [1928 rule(s)] done [0.03s]. |                              |         |       |        |   |  |  |  |
| Rules [#1928 association rules loa  | ided]                        |         |       |        |   |  |  |  |
| N° Antecedent   | Consequent                   | Support | Confi | Lift 🔻 |   |  |  |  |
| 714 own_telephone=none /\ existing_credits=one /\ other_payment_plans=none  | credit_history=existing_paid | 26.3    | 82.2  | 155.1  | ~ |  |  |  |
| 1474 existing_credits=one // housing=own // other_payment_plans=none // num_dependents=one  | credit_history=existing_paid | 25.3    | 80.3  | 151.5  |   |  |  |  |
| 779 existing_credits=one // housing=own // other_payment_plans=none   | credit_history=existing_paid | 28.7    | 80.2  | 151.3  |   |  |  |  |
| 1482 existing_credits=one // housing=own // other_payment_plans=none // foreign_worker=yes credit_history=existing_paid 27.1 79.7 15  |                              |         |       |        |   |  |  |  |
| 293 existing_credits=one /\ other_payment_plans=none  | credit_history=existing_paid | 41.5    | 79.5  | 150.0  | - |  |  |  |

In the upper part of the report, we can see the output of the BORGELT's program (version 4.31). In the lower part, the rules are enumerated. We can rank the rules according various numeric indicators by clicking on the column header. Because the component generates the rules with one item into the consequent, we obtain "only" 1928 rules.

#### 5 R (arules package)

The « arules » package (<u>http://cran.univ-lyon1.fr/web/packages/arules/index.html</u>) allows extracting association rules with R (<u>http://www.r-project.org/</u>). This is also a version of the BORGELT's program, with the same limitation. In comparison with Tanagra, we must explicitly prepare the dataset before. The attribute-value representation must be transformed into a transactional data format. The operation is easy... if we read carefully the documentation.

Loading the « arules » package. We use the library(.) command in order to load the package

```
#charger le package
library(arules)
```

**Data file importation and transformation**. We import the dataset with the **read.table(.)** command, **summary(.)** gives some indications about the data characteristics.

```
#charger le fichier de données
setwd("D:/DataMining/Databases_for_mining/comparison_TOW/association_rule")
german <- read.table(file="credit-german.txt",header=T,dec=".",sep="\t")
summary(german)</pre>
```

We cannot extract rules from a data.frame, we must transform the internal format in "transactions".

```
#transformer les données attributs-variables
#en données transactionnelles
german.trans <- as(german,"transactions")
summary(german.trans)</pre>
```

We have always 71 items. R gives indications about the density of the dataset. We obtain a large number of rules if the density of the database is high.

```
> summary(german.trans)
transactions as itemMatrix in sparse format with
1000 rows (elements/itemsets/transactions) and
71 columns (items) and a density of 0.2676056
most frequent items:
                             other_parties=none
     foreign_worker=yes
                                                      num_dependents=one other_payment_plans=none
                   963
                                            907
                                                                    845
                                                                                             814
            housing=own
                                        (Other)
                    713
                                          14758
element (itemset/transaction) length distribution:
sizes
 19
1000
  Min. 1st Qu. Median Mean 3rd Qu.
                                        Max.
          19
    19
                  19
                         19
                                19
                                         19
includes extended item information - examples:
                   labels
                              variables levels
1
      checking_status=<0 checking_status
                                              <0
  checking_status=>=200 checking_status >=200
2
3 checking_status=0<=X<200 checking_status 0<=X<200
includes extended transaction information - examples:
 transactionID
1
            1
2
             2
3
             3
```

Extraction of rules. The following instructions extract the association rules.

We obtain again the BORGELT's program output.

```
> #extraction des règles
> german.regles <- apriori(german.trans,parameter=
                 list(supp=0.25,conf=0.75,minlen=2,maxlen=10,target="rules"))
+
parameter specification:
confidence minval smax arem aval originalSupport support minlen maxlen target
                                                                                ext
      0.75 0.1 1 none FALSE
                                            TRUE 0.25 2 10 rules FALSE
algorithmic control:
filter tree heap memopt load sort verbose
   0.1 TRUE TRUE FALSE TRUE
                              2
                                   TRUE
apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)
                               (c) 1996-2004
                                              Christian Borgelt
set item appearances ... [0 item(s)] done [0.00s].
set transactions ...[71 item(s), 1000 transaction(s)] done [0.00s].
sorting and recoding items ... [31 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [1928 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
```

Some characteristics of the generated rule base are also available.

```
> summary(german.regles)
set of 1928 rules
rule length distribution (lhs + rhs):sizes
2 3 4 5 6
112 487 783 476 70
  Min. 1st Qu. Median
                      Mean 3rd Qu.
                                    Max.
 2.000 3.000 4.000 3.951 5.000 6.000
summary of quality measures:
  support
               confidence
                                  lift
Min. :0.2500 Min. :0.7500 Min. :0.8894
                             1st Qu.:0.9972
1st Qu.:0.2700 1st Qu.:0.8397
              Median :0.8906
Median :0.2990
                              Median :1.0121
Mean
     :0.3294
               Mean :0.8859
                              Mean :1.0437
              3rd Qu.:0.9485
3rd Qu.:0.3580
                              3rd Qu.:1.0475
Max. :0.8800 Max. :0.9937 Max. :1.5507
mining info:
       data ntransactions support confidence
german.trans 1000 0.25
                                0.75
```

**Visualization of the rules**. The inspect(.) command enables to visualize the details of rules. We show only the first 10 rules here.

| >  | #afficher les 10 premières règles trouvées           |    |                            |         |            |           |
|----|--|----|----------------------------|---------|------------|-----------|
| >  | <pre>inspect(german.regles[1:10])</pre>              |    |                            |         |            |           |
|    | lhs  |    | rhs                        | support | confidence | lift      |
| 1  | (checking_status=0<=X<200)                           | => | {foreign_worker=yes}       | 0.264   | 0.9814126  | 1.0191201 |
| 2  | (checking_status=<0)                                 | => | {foreign_worker=yes}       | 0.259   | 0.9452555  | 0.9815737 |
| 3  | (purpose=radio/tv)                                   | => | (num_dependents=one)       | 0.250   | 0.8928571  | 1.0566357 |
| 4  | (purpose=radio/tv)                                   | => | {foreign_worker=yes}       | 0.275   | 0.9821429  | 1.0198784 |
| 5  | <pre>(property_magnitude=real estate)</pre>          | => | {foreign_worker=yes}       | 0.262   | 0.9290780  | 0.9647747 |
| 6  | <pre>{credit_history=critical/other existing }</pre> | => | {other_payment_plans=none} | 0.251   | 0.8566553  | 1.0524021 |
| 7  | <pre>{credit_history=critical/other existing }</pre> | => | (other_parties=none)       | 0.268   | 0.9146758  | 1.0084628 |
| 8  | <pre>{credit_history=critical/other existing }</pre> | => | {foreign_worker=yes}       | 0.279   | 0.9522184  | 0.9888042 |
| 9  | (class=bad)  | => | (num_dependents=one)       | 0.254   | 0.8466667  | 1.0019724 |
| 10 | (class=bad)  | => | {other_parties=none}       | 0.272   | 0.9066667  | 0.9996325 |

We can also rank the rules according to a rule quality indicator. We show here the first 5 rules according to the LIFT criterion. We obtain the same rules as the A PRIORI PT component of Tanagra.

| > | #afficher les 5 règles avec  | $1\epsilon$ | e lift le + élevé                         |         |             |         |
|---|--|-------------|---|---------|-------------|---------|
| > | regles.triees <- sort(germa  | n.r         | regles,by="lift")                         |         |             |         |
| > | <pre>inspect(regles.triees[1:5])</pre>   |             |   |         |             |         |
|   | lhs  |             | rhs                                       | support | confidence  | lift    |
| 1 | <pre>{other_payment_plans=none,<br/>existing_credits=one,</pre>                |             |   |         |             |         |
| 2 | <pre>own_telephone=none;<br/>{other_payment_plans=none,<br/>housing=own,</pre> | =>          | {credit_nistory=existing paid}            | 0.263   | 0.8218750 1 | .550708 |
|   | existing_credits=one,<br>num_dependents=one)                                   | =>          | {credit_history=existing paid}            | 0.253   | 0.8031746 1 | .515424 |
| 3 | <pre>{other_payment_plans=none,<br/>housing=own,</pre>                         |             |   |         |             | 540505  |
| 4 | <pre>existing_credits=one; {other_payment_plans=none, housing=own,</pre>       | =>          | {credit_nistory=existing paid}            | 0.287   | 0.8016760 1 | .512596 |
| 5 | existing_credits=one,<br>foreign_worker=yes)<br>{other payment plans=none,     | =>          | <pre>(credit_history=existing paid)</pre> | 0.271   | 0.7970588 1 | .503885 |
|   | existing_credits=one)  | =>          | <pre>(credit_history=existing paid)</pre> | 0.415   | 0.7950192 1 | .500036 |

### 6 Orange

**Creation of a "schema" and data importation.** When we launch Orange, a new empty schema is available. We add the FILE component (DATA tab). We select the data file.



**Rule extraction and visualization**. We insert the ASSOCIATION RULES component (ASSOCIATE tab). We set the parameters of our analysis.



Then we add the ASSOCIATION RULES VIEWER component, we connect the components. We can now set the connection between FILE and ASSOCIATION RULES. The calculation is launched.



To see the rules, we click on the VIEW menu of the ASSOCIATION RULES VIEWER component.

| 📕 Qt Associati | on Rules Viewer          |                             |                 |        |            |           |                      |  |   |
|----------------|--------------------------|-----------------------------|-----------------|--------|------------|-----------|----------------------|--|---|
| Filter         |                          |                             |                 | - Rule | s          |           |                      |  | _ |
| Shown          | Support (H)<br>28% - 87% | Confidence (V)<br>75% - 99% | # Rules<br>1732 | •      | Support    |           | 🔽 Lift               | Strength                               |   |
| Selected       | 29% - 88%                | 75% - 99%                   | 1732            |        | Confidence |           | Leverage             | Coverage                               |   |
|                |                          |                             |                 | Sup    | p Conf     | Lift 🔻    | Rule                 |  | • |
|                |                          |                             |                 | 0.2    | 89 0.80    | 1 1.534   | credit_history=exist | ting paid class=good -> other_paymer   |   |
|                |                          |                             |                 | 0.3    | 19 0.76    | 9 1.534   | credit_history=exist | ting paid other_parties=none num_dep   |   |
|                |                          |                             |                 | 0.3    | 30 0.79    | 5 1.523   | credit_history=exist | ting paid other_parties=none num_der   |   |
|                |                          |                             |                 | 0.3    | 19 0.79    | 0 1.513   | credit_history=exist | ting paid other_parties=none num_dep   |   |
|                |                          |                             |                 | 0.3    | 65 0.79    | 0 1.513   | credit_history=exist | ting paid num_dependents=one -> otł    |   |
|                |                          |                             |                 | 0.3    | 50 0.75    | 8 1.512   | credit_history=exist | ting paid num_dependents=one -> otł    |   |
|                |                          |                             |                 | 0.3    | 56 0.75    | 4 1.505   | credit_history=exist | ting paid other_parties=none -> other_ |   |
|                |                          |                             |                 | 0.3    | 50 0.78    | 5 1.503   | credit_history=exist | ting paid num_dependents=one foreig    |   |
|                |                          |                             |                 | 0.3    | 70 0.78    | 4 1.502   | credit_history=exist | ting paid other_parties=none -> other_ |   |
|                |                          |                             |                 | 0.4    | 15 0.79    | 5 1.500   | other_payment_pla    | ans=none existing_credits=one -> crec  |   |
|                |                          |                             |                 | 0.4    | 15 0.78    | 3 1.500   | credit_history=exist | ting paid -> other_payment_plans=nor   |   |
|                |                          |                             |                 | 0.3    | 65 0.79    | 3 1.497   | other_payment_pla    | ans=none existing_credits=one num_de   |   |
|                |                          |                             |                 | 0.3    | 50 0.76    | 1 1.492   | other_payment_pla    | ans=none existing_credits=one num_de   |   |
|                |                          |                             |                 | 0.3    | 96 0.79    | 0 1.491   | other_payment_pla    | ans=none existing_credits=one foreign  |   |
|                |                          |                             |                 | 0.3    | 19 0.82    | 0 1.491   | credit history=exist | ting paid other parties=none other pa  |   |
|                |                          |                             |                 | 0.3    | 19 0.79    | 2 1 / 91  | oradit history-avist | ting paid other parties-none other pa  | • |
|                |                          |                             |                 |        |            |           |                      | Send rules automatically               |   |
| Zoom           | Show All                 | No Zoom                     | Unselect        |        |            | Save Rule | es                   | Send Rules                             |   |

The visualization window is really original. We can graphically select a group of rules according to a range of the value of two criteria. Various criteria can be used; we can also rank the rules here.

| 📕 Qt Association            | n Rules Viewer                        |  |                      |   |                             |                |           |                               |  |            |
|-----------------------------|---------------------------------------|--|----------------------|---|-----------------------------|----------------|-----------|-------------------------------|--|------------|
| Filter<br>Shown<br>Selected | Support (H)<br>28% - 87%<br>58% - 61% | Confidence (V)<br>75% - 99%<br>96% - 97% | # Rules<br>1732<br>3 |   | Rules —<br>I Supp<br>I Conf | oort<br>idence |           | ✓ Lift<br>Leverage            | ☐ Strength<br>☐ Coverage                 |            |
|                             |                                       |  |                      |   | Supp                        | Lonf           | Lift      | Hule                          |  |            |
|                             |                                       |  |                      |   | 0.601                       | 0.362          | 0.333     | evisiting credits=one () fore | iooa -> Toreign_worke<br>aign_worker=ues | er=yes     |
|                             |                                       |  |                      |   | 0.586                       | 0.965          | 1.002     | housing=own num depende       | ents=one -> foreian v                    | worker=ves |
|                             |                                       |  |                      | ' |                             |                |           |                               |  |            |
|                             |                                       |  |                      |   |                             |                |           | l⊄ Sendiru                    | ules automatically                       |            |
| Zoom                        | Show All                              | No Zoom                                  | Unselect             |   |                             | Sa             | ave Rules |                               | Send Rules                               |            |

## 7 RapidMiner

**Creating a new operator tree**. With RapidMiner, it is more convenient to define before all the sequences of operations before starting the computation. The first operator is **CSVEXAMPLESOURCE**, we set the file name and the column separator.

| 😵 RapidMiner@FUJITSU (assoc rule on german.>  | ml)  |                                     |
|---|--|-------------------------------------|
| <u>F</u> ile <u>E</u> dit ⊻iew <u>P</u> rocess <u>T</u> ools <u>H</u> elp   |  |                                     |
| 🞦 📁 📗 🖶 🗞 🛛 🛯   | 🍉 🔎 🍑 🕨 🖋  | 🚨 💓 😨                               |
| Coperator Tree  | Parameters 🕞 XML 📄 Com   | ment 🥻 🤷 New Operator               |
| ⊟- ■ Root   | filename   | Aassociation_rule\credit-german.txt |
|   | read_attribute_names   |                                     |
|   | label_name   |                                     |
| Nominal2Binominal   | label_column   | 0                                   |
| FPGrowth  | id_name  |                                     |
| FPGrowth  | id_column  | 0                                   |
| AssociationRuleGenerator  | weight_name  |                                     |
|   | weight_column  | 0                                   |
|   | sample_ratio   | 1.0                                 |
|   | sample_size  | -1                                  |
|   | datamanagement   | double_array 💌                      |
|   | column_separators  | u 🔶 📼                               |
| [disc_amount = up_2000, personal_status = male sin;<br>(confidence: 0.753)<br>[class = good]> [housing = own] (confidence: 0.753)<br>[job = skilled, personal_status = male single]> [housing =<br>[num_dependents = one, class = good]> [housing =<br>[other_parties = none, age = 30<=X<55]> [num_dependents = none, age = 30<=X<55]> [num_dependents = none, housing =<br>(confidence: 0.753)<br>[housing = own]> [foreign_worker = yes, other_parties | jiej> [num_dependents = one]<br>ing = own] (confidence: 0.753)<br>own] (confidence: 0.753)<br>endents = one] (confidence: 0.753)<br>own, job = skilled]> [class = good]<br>s = none, num_dependents = one] | Max: 1.1 GB<br>Total: 1.1 GB        |
| •   |  | 10:50:18 PM                         |

RapidMiner cannot create association rules from an attribute-value dataset. We must recode the variable into a set of binary columns with the **NOMINAL2BINOMIAL** component.

| 😵 RapidMiner@FUJITSU (assoc rule on german.                                       | .xml)                      |                            |             |
|---|----------------------------|----------------------------|-------------|
| <u>F</u> ile <u>E</u> dit <u>V</u> iew <u>P</u> rocess <u>T</u> ools <u>H</u> elp |                            |                            |             |
| 🎦 📁 🔡 🖶 🗞 🔗 🛯   | 🍋 🔎 🌘 🖉 🧭                  | 2                          | 1           |
| Coperator Tree  | Parameters 🕞 XML 📄 Comme   | nt 🥻 🐿 New Operator        |             |
| E Coot  | return_preprocessing_model |                            |             |
|   | create_view                |                            |             |
|   | transform_binominal        |                            | -           |
| Nominal2Binominal   | use_underscore_in_name     |                            |             |
| FPGrowth  |                            |                            |             |
| AssociationRuleGenerator<br>AssociationRuleGenerator                              |                            |                            |             |
|   |                            | ax: 1.1 CB<br>oral: 1.1 GD |             |
| Θ   |                            |                            | 10:57:46 PM |

The extraction is carried out in 2 steps. First, with the **FPGROWTH** component, we generate the frequent itemsets. The settings must be defined carefully.

| RapidMiner@FUJITSU (assoc rule on german.)  | xml)                        |                      |
|---|-----------------------------|----------------------|
| <u>F</u> ile <u>E</u> dit <u>V</u> iew <u>P</u> rocess <u>T</u> ools <u>H</u> elp |                             |                      |
| 🞦 📁 📕 🖶 🗞 🛛 🛯   | 🐿 🔎 🍯 🔰 💜                   | 🚨 💓 😨                |
| Coperator Tree  | Parameters 🕞 XML 📄 Comm     | ent 🏻 🦜 New Operator |
| ⊟- ■ _ Root   | keep_example_set            |                      |
| Process   | find_min_number_of_itemsets |                      |
| CSVExampleSource  | min_number_of_itemsets      | 100                  |
| - Anninal2Binominal   | min_support                 | 0.25 🧲               |
| EPGrowth  | max_items                   | -1                   |
| FPGrowth  | must_contain                |                      |
| AssociationRuleGenerator  |                             |                      |
|   |                             |                      |
|   |                             |                      |
|   |                             |                      |
|   |                             |                      |
|   |                             |                      |
|   |                             | Max: 1.1 GB          |
|   |                             |                      |
| Θ   |                             | 11:02:02 PM          |

Then, with the **ASSOCIATIONRULEGENERATOR**, we generate the rules from the itemsets.

| RapidMiner@FUJITSU (assoc rule on german.)           | aml)                    |                              |             |
|--|-------------------------|------------------------------|-------------|
| <u>File Edit View Process Tools H</u> elp            |                         | -                            | im da vila  |
| 🞦 🍋 📕 🖶 📚 🛛 🖉  | 🎽 🖉 🦉 🕨 📕 💜             | 2                            |             |
| Coperator Tree                                       | Parameters 🕞 XML 📄 Comm | ent 🧯 New Operator           |             |
| ⊟- ■<= Root<br>Process                               | keep_frequent_item_sets |                              |             |
|  | min_confidence          | 0.75 🧲                       |             |
| CSVExampleSource                                     | gain_theta              | 2.0                          |             |
| Nominal2Binominal                                    | laplace_k               | 1.0                          |             |
| FPGrowth FPGrowth                                    |                         |                              |             |
| AssociationRuleGenerator<br>AssociationRuleGenerator |                         |                              |             |
|  |                         |                              |             |
|  |                         | Max: 1.1 GB<br>Total: 1.1 GD |             |
| Θ  |                         | · · · · · ·                  | 11:04:57 PM |

**Launching the computation and visualization of rules**. We launch the calculations by clicking on the PLAY menu into the toolbar.

| RapidMiner@FUJITSU (assoc rule on german.xml)  |                                   |                   |                 |             |                  |               |         | X      |         |        |       |        |        |               |
|--|-----------------------------------|-------------------|-----------------|-------------|------------------|---------------|---------|--------|---------|--------|-------|--------|--------|---------------|
| <u>F</u> ile <u>E</u> dit <u>V</u> iew <u>P</u> rocess <u>T</u>  | Eile Edit View Process Tools Help |                   |                 |             |                  |               |         |        |         |        |       |        |        |               |
| 🞦 📁 📕 🖶 🔌  | 5                                 | (N) 🔒             | æ 🍯             |             |                  | 2             |         |        |         |        |       | 200    | 2 1    | Į             |
| 💡 Association Rules  |                                   |                   |                 |             |                  |               |         |        |         |        |       |        |        |               |
| 💿 Table View 🔵 Graph Viev  | w 🔘 Tex                           | t View            |                 |             |                  |               |         |        |         |        |       | V      |        |               |
| Conjunction Type:  | No.                               | Pi                | remises         |             | Conclu           | sion          | Support | Confid | LaPlace | Gain   | p-s   | Lift 🔻 | Conv   |               |
| And  | 958                               | other_payment_p   | olans = none,   | existing_   | credit_history = | existing paid | 0.263   | 0.822  | 0.957   | -0.377 | 0.093 | 1.551  | 2.639  |               |
|  | 794                               | num_dependent     | s = one, class  | = good, (   | other_payment    | _plans = nor  | 0.253   | 0.808  | 0.954   | -0.373 | 0.090 | 1.548  | 2.494  |               |
| Conclusions:   | 273                               | other_parties = n | one, num_de     | pendents    | foreign_worker   | = yes, other_ | 0.319   | 0.769  | 0.932   | -0.511 | 0.111 | 1.534  | 2.157  |               |
| foreign_worker = yes 🛛 🔼   | 702                               | class = good, cre | dit_history = e | existing p: | other_payment    | _plans = nor  | 0.289   | 0.801  | 0.947   | -0.433 | 0.101 | 1.534  | 2.397  |               |
|  | 662                               | own_telephone =   | none, credit    | history =   | other_payment    | _plans = nor  | 0.263   | 0.797  | 0.950   | -0.397 | 0.091 | 1.527  | 2.354  |               |
|  | 660                               | other_parties = n | one, class = g  | jood, crea  | other_payment    | _plans = nor  | 0.255   | 0.797  | 0.951   | -0.385 | 0.088 | 1.527  | 2.353  |               |
| Min. Confidence:   | 640                               | other_parties = n | one, num_dej    | pendents    | other_payment    | _plans = nor  | 0.330   | 0.795  | 0.940   | -0.500 | 0.113 | 1.523  | 2.334  |               |
|  | 609                               | foreign_worker =  | yes, class = g  | jood, crea  | other_payment    | _plans = nor  | 0.272   | 0.793  | 0.947   | -0.414 | 0.093 | 1.519  | 2.309  | $\overline{}$ |
|  |                                   |                   |                 |             |                  |               |         |        |         |        |       | (      | Save   |               |
| other_payment_plans = nonel (contidence: U.75b)<br>[class = good, credit_history = existing paid]> [foreign_worker = yes, other_parties = none, num_dependents =<br>one) (confidence: 0.756)<br>[other_payment_plans = none) (confidence: 0.756) |                                   |                   |                 |             |                  |               |         |        |         |        |       |        |        |               |
| other_paymen_plans = nonej (contidence: U.756)<br>[age = 30<=X<55]> [num_dependents = one] (confidence: 0.756)<br>2886 other rules<br>(created by AssociationRuleGenerator)<br>P Nov 16, 2008 11:06:40 PM: [NOTE] Process finished successfully  |                                   |                   |                 |             |                  |               |         |        |         |        |       |        |        |               |
| Θ  |                                   |                   |                 |             |                  |               |         |        |         |        |       | 11     | :08:35 | ΡM            |

We obtain 2986 rules, the same as the A PRIORI component of Tanagra. Various indicators enable to rank rules. Below, we sort the rules according to the LIFT criterion.

| 😵 RapidMiner@FUJITSU (asso  | oc rule on german.xml)   |            |  |  |  |  |
|---|--|------------|--|--|--|--|
| <u>F</u> ile <u>E</u> dit <u>V</u> iew <u>P</u> rocess <u>T</u> o   | ools <u>H</u> elp  |            |  |  |  |  |
| 睯 🍯 📕 😓 📚   | 🔊 🔍 🐂 🥦 🕨 🔳 💉 💄  | 2 🖷        |  |  |  |  |
| 💡 Association Rules   |  |            |  |  |  |  |
| Table View O Graph View   | C Text View  |            |  |  |  |  |
| Conjunction Type:   | No. Premises Conclusion SuppConfLaPI Gain p-s Lit  | ft⊽ Con    |  |  |  |  |
| And   | 243 other_parties = none, other_payment_plans = none, checking_status class = good 0.290 0.927 0.982 -0.33 0.071 1.                | .324 4.080 |  |  |  |  |
|   | 241 foreign_worker = yes, other_parties = none, other_payment_plans = r class = good 0.280 0.924 0.982 -0.32 0.068 1.              | .32( 3.951 |  |  |  |  |
| Conclusions:  | 241 other_parties = none, num_dependents = one, other_payment_plans class = good 0.250 0.923 0.983 -0.29 0.060 1                   | .311 3.871 |  |  |  |  |
| foreign_worker = yes  | 237 other_payment_plans = none, checking_status = no checking class = good 0.303 0.918 0.988 -0.35 0.072 1.                        | .311 3.667 |  |  |  |  |
| other_parties = none  | 233 num_dependents = one, other_payment_plans = none, checking_sta class = good 0.261 0.916 0.981 -0.30 0.062 1.                   | .301 3.561 |  |  |  |  |
| num_dependents = one  | 231 foreign_worker = yes, other_payment_plans = none, checking_status class = good 0.291 0.916 0.98C -0.34 0.068 1.                | .301 3.53( |  |  |  |  |
| other_payment_plans = none  | 228 foreign_worker = yes, num_dependents = one, other_payment_plans class = good 0.253 0.913 0.981 -0.30 0.05§ 1.                  | .30( 3.461 |  |  |  |  |
| nousing = own   | 205 other_parties = none, housing = own, checking_status = no checking_class = good_0.259_0.896_0.9770.31_0.057_1.                 | .28( 2.89( |  |  |  |  |
| existing credits = one  | 203 housing = own, checking_status = no checking class = good 0.272 0.896 0.976 -0.33 0.056 1.                                     | .271 2.851 |  |  |  |  |
| credit history = existing paid  | 201 foreign_worker = yes, other_parties = none, housing = own, checking class = good 0.251 0.893 0.977 -0.31 0.054 1.              | .27€ 2.81€ |  |  |  |  |
| Min Confidence:   | 198 foreign_worker = yes, housing = own, checking_status = no checking   class = good   0.263   0.892   0.975   -0.32   0.057   1. | .27+ 2.76  |  |  |  |  |
| Mini: Confidence.   |  |            |  |  |  |  |
|   |  |            |  |  |  |  |
|   |  | Save       |  |  |  |  |
| other_payment_plans = none] (confidence: U./56)<br>(class = good, credit_history = existing paid]> [foreign_worker = yes, other_parties = none, num_dependents =<br>one] (confidence: 0.756)<br>[other_payment plans = none] (confidence: 0.756)                        |  |            |  |  |  |  |
| Outrier_parine = nonej (connidence: 0.756)         [age = 30 <= x < 55] ···> [num_dependents = one] (confidence: 0.756)         2886 other rules         (created by AssociationRuleGenerator)         P Nov 16, 2008 11:06:40 PM: [NOTE] Process finished successfully |  |            |  |  |  |  |
| Θ   |  | 1:16:47 PM |  |  |  |  |

Another option is available. We can filter out the rules according to the presence or absence of an item or a set of items. This is very useful.

#### 8 Knime

A double data preparation is necessary for KNIME before launching the learning algorithm: a coding o/1 of the attribute-value dataset, followed by a transformation into transactional data.

**Workflow creation and data importation**. We create a new project by clicking on the FILE / NEW menu. We choose a KNIME project and we specify the name of the project.

| File Edit View Se                      | arch Node Help  |  |
|--|---|--|
| 📸 New 🗕 🚬                              | A New   | 🗙 🔊 🖉 🍇 🗄 🐨  |
| 🕌 Save<br>🎼 Save All                   | Select a wizard<br>This wizard creates a new KNIME workflow project.  |  |
| Switch Workspac                        | _/  | New KNIME Project Wizard   |
| Preferences<br>Update KNIME            | Wizards:<br>type filter text  | Create a new Knime workflow project.                             |
| Exit<br>Mining<br>Exit<br>Meta<br>Misc | New KNIME Project  File Folder Folder Votile Project File Folder Konstanz Information Miner New KNIME Project | Name of the project to create: Association Rules - German credit |
|  | Log<br>WARN<br>WARN   | <pre> Back Next &gt; Finish Cancel </pre>                        |
| L                                      |   | 18W OF 44W   |

An empty workflow appears. We insert the file access component (FILE READER). We select our data file (CONFIGURE menu). We click on the EXECUTE menu in order to load the dataset.

| <b>▲</b> KNIME  |  |
|---|--|
| File Edit View Search Node Help                             |  |
| i 📬 ▪ 🔛 🔞 i 🛷 i ½ × 🖓 × i 🍫 i 💝 🏷   100%                    | 🗸 📾 늘 🗗 🔘 🕲 🕲 🖉 i 📴 🗸  |
| 🔊 Node Repository 🛛 🗖 🗖 🔺 Association Rules - German credit |  |
| $\bigtriangledown$  | Dialog - File Reader (#1)  |
| V File Reader   | File   |
|   | Settings General Node Settings   |
|   | Enter ASCII data file location: (press 'Enter' to update preview)  |
| Read A  |  |
| ARFF Reader Node 1  | valid URL: ses for mining/comparison TOW/association rule/credit-german.txt V Brov   |
| File Reader   | -Paris Sattings  |
| Table Reader  | Dask Jettings  |
| Write     Execute and open view                             | Column delimiter: <cab> Advanced</cab>   |
| 🗈 🗁 Artificial Data 🧰 Cancel                                | Image: Version of the second se |
| Cache   | Java-style comments Single line comment:   |
| Database     Database     Node name and description         |  |
| B Q Data Views  | - Draviau  |
| E Statistics  | Click column beader to change column properties (* = name(type user settings)  |
| Mining     Paste  |  |
|   | Row ID S checkin S disc_du S credit_history S purpose :  |
| 🔲 🕓 Redo  | Row2 0<=X<200 up_2_years existing paid radio/tv u  |
| KNIM 💥 Delete   |  |
| WAR Q Data Outport 0: File Table                            | WARNING: suggested settings are based on a partial file analysis only! Please verify.  |
| WAR File Reader   |  |
|   |  |
|   |  |
|   | 32M of 78M   |

**First coding step**. First we must recode the attribute-value dataset into a binary dataset. We use the ONE2MANY component (DATA MANIPULATION / COLUMN). We connect this component to the

previous one. We add an INTERACTIVE TABLE component (DATA VIEWS), that we connect to ONE2MANY, in order to view the new dataset.

| A KNIME                            |   |          |
|------------------------------------|---|----------|
| File Edit View Search Node Help    |   |          |
| : = + - 🗆 🖻 : 🖉 : 🖣 - 🗏 - : *5 : 4 |   |          |
|                                    |   |          |
| 🔊 Workflow Projects 🛛 👘 🗖 🔊 *Asso  | iation Rules - German credit 🗴 🔊 Règles d'Association |          |
| A Node Repository 22               |   |          |
| File                               | Reader One2Many Interactive Table                     |          |
|                                    |   |          |
| <u> </u>                           |   |          |
|                                    |   |          |
| String To Number 🔼 N               | de 1   F O   Node 4                                   |          |
| 🕀 💶 Row                            | A Interactive Table (#4) - Table View (1000 x 90)     |          |
| 🕀 🛗 Matrix                         | File Hilite Navigation View Output                    |          |
| 🖃 🔍 Data Views                     | Row ID S for S class T < T 0 T por T > T lo T         |          |
| 💷 🛬 Property                       |   |          |
| 夏 Box Plot                         |   |          |
| Conditional Box Plot               | Row3 yes good 0 0 1 0 1 0                             | 0        |
| Histogram                          | Row4 yes good 1 0 0 0 1                               | 0        |
| Histogram (interactive)            | Row5 yes bad 1 0 0 0 0                                | 1        |
| Interactive Table                  | Row6 yes good 0 0 1 0 0 1                             | 0        |
| V Line Plot                        | Row7 yes good 0 0 1 0 0 0                             | 1        |
| Parallel Coordinates               | e Row8 yes good 0 1 0 0 1                             | 0        |
| Pie chart KNIME Cor                | Row9 yes good 0 0 1 0 1 0                             | 0        |
| Pie chart (interactive)            | Row10 yes bad 0 1 0 0 0 1                             | 0        |
| Rule2DPlotter                      | Row11 yes bad 0 1 0 0 1 0                             | 0        |
| 「対応 Scatter Matrix                 | Row12 yes pad 1 U U U U 1                             |          |
| Scatter Plot                       | Row14 yes bad 1 0 0 0 0 0                             |          |
|                                    |   | <u>+</u> |
|                                    |   | 1 or 99M |

There are 90 columns now. To the previous 19 variables are added 71 binary variables.

**Second coding step**. The first transformation is not enough. We must go through an internal format specific using the BITVECTOR component (DATA MANIPULATION / COLUMN). We connect it to ONE2MANY. We click on the CONFIGURE menu, we select only the binary variables.

| A KNIME                         |                     |                                   |                 |  |
|---------------------------------|---------------------|-----------------------------------|-----------------|--|
| File Edit View Search Node Help |                     |                                   |                 |  |
| 1 📬 • 🗐 🕼 1 🛷 1 🖢 - 🖄 •         | : 🍋 : 🥔 🌭 🛛 100     | % 🗸 🚮                             |                 | 2 🔘 🖉 : 🕞 •                                |
|                                 |                     | Courses and it                    |                 |  |
|                                 | Association Rules - | German credic                     |                 | Sociation                                  |
| 🔊 Node Repository 🔀 📃 🗖         |                     |                                   |                 | Dialog - Diffector Generator (#5)          |
| ~                               | File Reader         | One2Many                          | Interactive Tal | File                                       |
| <b>~</b>                        |                     | -> <mark>=</mark> >- <sub>1</sub> | ⊳ 🔳             | Default Settings General Node Settings     |
|                                 |                     |                                   |                 | Numeric input (many columns)               |
| 😟 🗁 Artificial Data 🛛 🗛         | Node 1              |                                   | Node 4          | Threshold: 1 📚                             |
|                                 |                     | Node 3                            | waster Caparat  | Use percentage of the mean:                |
| 🕀 😫 Database                    |                     | DI                                |                 |  |
| Data Manipulation               |                     |                                   |                 | Percentage: 100 📚                          |
|                                 |                     |                                   |                 | Parse bitvectors from strings (one column) |
| Bitvector Generator             |                     |                                   |                 |  |
|                                 |                     |                                   | NUCES           | String column to be parsed St class        |
| GroupBy                         |                     |                                   |                 |  |
| 🕂 🕂 HiLite Filter               |                     |                                   |                 |  |
| Numeric Row Splitter            |                     |                                   |                 | Kind of string representation: BIT 🕑       |
| Partitioning                    | 📃 Console           |                                   |                 |  |
| Reference Row Filter            | KNIME Console       |                                   |                 | Replace column(s)                          |
|                                 | WARN One2 Ma        | any                               | Model warn      |  |
|                                 | WARN One2 Ma        | any                               | Model warn      |  |
| -⊃⊄ Shuffle                     |                     |                                   |                 | OK Apply Cancel                            |
| 🔤 🔠 Sorter                      | 1                   |                                   |                 |  |
| i 📅 🗰 Kārkuit.                  |                     |                                   |                 |  |
|                                 |                     |                                   |                 | 62M of 99M                                 |

By clicking on the EXECUTE AND OPEN VIEW menu, we obtain a description of the generated transactional dataset.

| Bitvector Generator (#5   | i) 🔳 🗖 🔀     |
|---------------------------|--------------|
| File                      |              |
| BitVector Generator       | Information: |
| Number of processed rows: | 1000         |
| Total number of 0s:       | 52000        |
| Total number of 1s:       | 19000        |
| Ratio of 1s to 0s:        | 0.3653       |

To obtain the information of density of R, we make 19000/(19000+52000) = 26.76%.

**Extraction of rules**. We use the ASSOCIATION RULE LEARNER component in order to extract rules. We set the following parameters.



The INTERACTIVE TABLE component enables to visualize the rules.

| A KNIME                         |  |
|---------------------------------|--|
| File Edit View Search Node Help |  |
| : 🐟 - 🖂 🗈 🖉 : 🗛 - 🔍 -           |  |
|                                 |  |
| 🔊 Workflow Projects 🖾 👘 🗆       | A *Association Rules - German credit X A Règles d'Association        |
| 🔺 Node Repository 🖾 📃 🗖         |  |
|                                 | File Reader One2Many Interactive Table                               |
|                                 |  |
|                                 |  |
| Interactive Table               | Node 1 Node 4 Association Interactive Table                          |
| V Line Plot                     | Node 3 Bitvector GeneratorRule Learner                               |
| 📲 Parallel Coordinates          |  |
|                                 |  |
|                                 |  |
| 📃 Rule2DPlotter                 | Node 5 Node 7 Node 9   |
| 기분 Scatter Matrix               |  |
| Scatter Plot                    |  |
| E Statistics                    |  |
| 🖨 🎯 Mining                      |  |
| 🖃 🗁 Association Rules           |  |
| Association Rule Learne         |  |
| 👬 Bitvector Generator           | KIME Coppes  |
| 🗈 🗁 Bayes                       | WARN One2Many Model warning message: Dunlicate nossible values found |
| 🕀 🧁 Clustering 🔤                | WARN One2Many Model warning message: Duplicate possible values found |
| 🕀 🗁 Rule Induction              | WARN Association Rule Learner Configure failed: Set the column with  |
| 🗈 🗁 Neural Network 🔍            |  |
|                                 |  |
|                                 | 73M of 100M  |

KNIME extracts only the rules with on item into the consequent. 1928 rules are generated. The visualization tool has not interactive functionalities (ranking or filtering rules).

| <b>▲</b> Interactive T | 🛿 Interactive Table (#9) - Table View (1928 x 14) |           |                          |           |                   |            |             |             |             |
|------------------------|---|-----------|--------------------------|-----------|-------------------|------------|-------------|-------------|-------------|
| File Hilite Naviga     | ile Hillte Navigation View Output                 |           |                          |           |                   |            |             |             |             |
| Row ID                 | D Support   | D Confide | S Consequent             | S implies | S Item0           | S Item1    | S Item2     | S Item3     |             |
| 📰 rule0 🛛 🚺            | 0.259   | 0.945     | yes_foreign_worker       | <         | <0_checking_sta   | ?          | ?           | ?           | 3           |
| rule1                  | 0.264   | 0.981     | yes_foreign_worker       | <         | 0<=X<200_chec     | ?          | ?           | ?           | 7           |
| rule2                  | 0.373   | 0.947     | none_other_parties       | <         | no checking_chec  | ?          | ?           | ?           | 7           |
| rule3                  | 0.313   | 0.948     | none_other_parties       | <         | no checking_chec  | none_other | ?           | ?           | 7           |
| rule4                  | 0.313   | 0.839     | none_other_payment_plans | <         | no checking_chec  | none_other | ?           | ?           | 7           |
| rule5                  | 0.271   | 0.951     | none_other_parties       | <         | no checking_chec  | none_other | one_num_d   | ?           | 7           |
| rule6                  | 0.271   | 0.86      | none_other_payment_plans | <         | no checking_chec  | none_other | one_num_d   | ?           | 7           |
| rule7                  | 0.271   | 0.866     | one_num_dependents       | <         | no checking_chec  | none_other | none_other  | ?           | 7           |
| rule8                  | 0.264   | 0.953     | none_other_parties       | <         | no checking_chec  | none_other | one_num_d   | yes_foreign | 7           |
| rule9                  | 0.264   | 0.86      | none_other_payment_plans | <         | no checking_chec  | none_other | one_num_d   | yes_foreign | 7           |
| rule10                 | 0.264   | 0.871     | one_num_dependents       | <         | no checking_chec  | none_other | none_other  | yes_foreign | 7           |
| rule11                 | 0.264   | 0.974     | yes_foreign_worker       | <         | no checking_chec  | none_other | none_other  | one_num_d   | 7           |
| rule12                 | 0.25  | 0.958     | none_other_parties       | <         | no checking_chec  | none_other | one_num_d   | good_class  | 7           |
| rule13                 | 0.25  | 0.893     | none_other_payment_plans | <         | no checking_chec  | none_other | one_num_d   | good_class  | 7           |
| rule14                 | 0.25  | 0.862     | one_num_dependents       | <         | no checking_chec  | none_other | none_other  | good_class  | 7           |
| rule15                 | 0.25  | 0.923     | good_class               | <         | no checking_chec  | none_other | none_other  | one_num_d   | 7           |
| rule16                 | 0.303   | 0.953     | none_other_parties       | <         | no checking_chec  | none_other | yes_foreign | ?           | 7           |
| rule17                 | 0.303   | 0.837     | none_other_payment_plans | <         | no checking_chec  | none_other | yes_foreign | ?           | 7           |
| rule18                 | 0.303   | 0.968     | yes_foreign_worker       | <         | no checking_chec  | none_other | none_other  | ?           | 7           |
| rule19                 | 0.28  | 0.962     | none_other_parties       | <         | no checking_chec  | none_other | yes_foreign | good_class  | 7           |
| rule20                 | 0.28  | 0.875     | none_other_payment_plans | <         | no checking_chec  | none_other | yes_foreign | good_class  | 7           |
| rule21                 | 0.28  | 0.966     | yes_foreign_worker       | <         | no checking_chec  | none_other | none_other  | good_class  | 7           |
| rule22                 | 0.28  | 0.924     | good_class               | <         | no checking_chec  | none_other | none_other  | yes_foreign | 7           |
| nula?3                 | 0.20  | n 057     | none other parties       | /         | no checking, chec | none other | good class  | 2           | E <u></u> ≦ |

#### 9 Weka

Like RapidMiner, it is more convenient to define the whole treatment before launching the computation with Weka. We use the Knowledge Flow in this tutorial.

**Specifying the operations**. When we launch Weka, we click on the APPLICATIONS / KNOWLEDGEFLOW menu. First, we must insert the data access component CSV LOADER (DATASOURCES tab). We select the data file.



We add the APRIORI component (ASSOCIATIONS tab). We connect the previous component to this one. We click on the CONFIGURE menu in order to define the parameters.

| 📚 Weka 3.5.6 - KnowledgeFlow                     |                      |                                   |
|--|----------------------|-----------------------------------|
| Program Applications Tools Visualization Windows | <b>4</b>             |                                   |
| 🛎 KnowledgeFlow                                  | About                |                                   |
| DataSources DataSinks Filters Classifiers Cl     | Class implementing   | g an Apriori-type algorithm. More |
| Associations                                     | -                    | Capabilities                      |
|  | car                  | False                             |
| Filtered<br>Apriori Associat                     | classIndex           | -1                                |
|  | delta                | 0.05                              |
| C Knowledge Flow Layout                          | lowerBoundMinSupport | 0.25                              |
|  | metricType           | Confidence                        |
| data3et  | minMetric            | 0.75                              |
| CSULoader  | numRules             | 10000                             |
| Apriori  | outputItemSets       | False 💌                           |
| <  | removeAllMissingCols | False 💌                           |
| Status   | significanceLevel    | -1.0                              |
| OK   | upperBoundMinSupport | 1.0                               |
|  | verbose              | False                             |

Setting NUMRULES to 1000 removes the restriction on the number of generated rules. We add then the TEXT VIEWER component in order to visualize the rules.

We launch the calculations by clicking on the START LOADING menu of... the CSV LOADER component into the knowledge flow.



To visualize the results, we click on the SHOW RESULTS menu of the TEXTVIEWER component. Like the A PRIORI component of Tanagra, we obtain 2986 rules.

| 🕌 Text Viewer             |   |
|---------------------------|---|
| Result list               | Text  |
| 07:31:18 - Model: Apriori | Minimum support: 0.25 (250 instances)<br>Minimum metric <confidence>: 0.75<br/>Number of cycles performed: 15</confidence>  |
|                           | Generated sets of large itemsets:   |
|                           | Size of set of large itemsets L(1): 31  |
|                           | Size of set of large itemsets L(2): 162   |
|                           | Size of set of large itemsets L(3): 345   |
|                           | Size of set of large itemsets L(4): 334   |
|                           | Size of set of large itemsets L(5): 140   |
|                           | Size of set of large itemsets L(6): 16  |
|                           | Best rules found:   |
|                           | <ol> <li>other_parties=none property_magnitude=car 318 ==&gt; foreign_worker=yes 316 conf:(0.99)</li> <li>other_parties=none property_magnitude=car num_dependents=one 280 ==&gt; foreign_worker=yes</li> </ol> |
|                           | 3. property_magnitude=car housing=own 271 ==> foreign_worker=yes 269 conf:(0.99)<br>4. property_magnitude=car other_payment_plans=none 268 ==> foreign_worker=yes 266 conf:()                                   |
|                           |   |

## 10 Conclusion

All software presented in this tutorial can extract association rules from a data file in an "individuals x variables" format. For some of them, a data preparation is required before to produce a transactional data format. This is not always obvious, especially when the software is not well documented. I admit to having groping a bit. But finally, once the data are correctly generated, the software produced similar results. This is what matters.