

This document outlines the measures to assess association rules proposed by the **A PRIORI MR** and **SPV ASSOC RULE** components. They come from studies reported in some publications of A. Morineau and R. Rakotomalala (see <http://eric.univ-lyon2.fr/~ricco/publications.html>).

A measure characterizes the relevance of a rule. It can be used to rank them. It should also help to discern those that are "significantly interesting" from those who are irrelevant. This last point is totally prospective. There is no really satisfactory solution at this time.

1 Data table

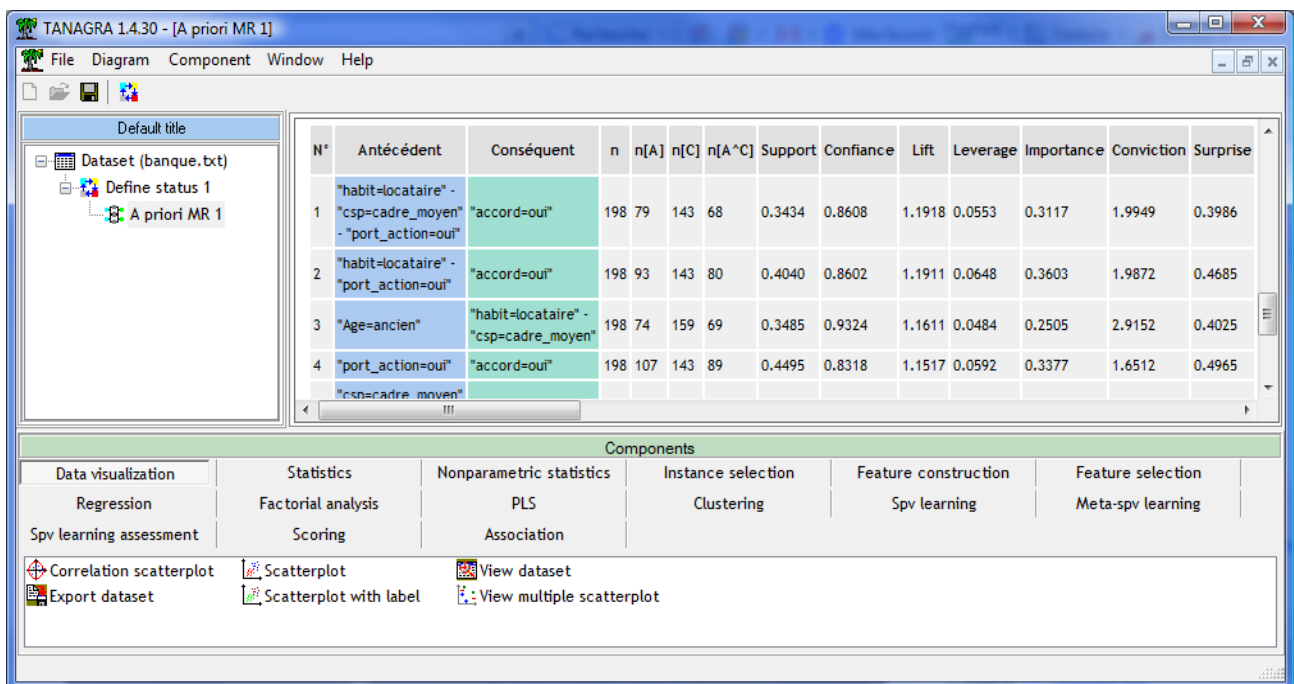
First, here are some indications about the notations used. A rule is composed of an antecedent and a consequent, which are composed of a list of items. We can summarize the number of individuals covered by a rule in the following table:

| | Antecedent | Not (Antecedent) | Sum |
|------------------|------------|------------------|-------|
| Consequent | n_{ac} | | n_c |
| Not (Consequent) | | | |
| Sum | n_a | | n |

With:

- n is the number of examples in the whole dataset;
- n_a is the number of examples covered by the antecedent of the rule;
- n_c is the number of examples covered by the consequent of the rule;
- n_{ac} is the number of examples covered by the rule i.e. both the antecedent and the consequent.

In the screenshot below, we show the results of the A PRIORI MR component on a dataset:



For the rule n°4: « IF port action = oui THEN accord = oui », « port action = oui » is the antecedent, « accord = oui » the consequent. We obtain the following values:

| | | | |
|------------------|------------|------------------|-----|
| | Antecedent | Not (Antecedent) | Sum |
| Consequent | 89 | | 143 |
| Not (Consequent) | | | |
| Sum | 107 | | 198 |

The measures supplied by Tanagra are computed from these values.

2 Classical measures for rule assessment

We name "classical" these measures for differentiating them to the indicators based on the "test value" concept that we will outline in the next section.

| Measure | Formula | Value |
|------------|---|--|
| Support | $\frac{n_{ac}}{n}$ | $\frac{89}{198} = 0.4495$ |
| Confiance | $\frac{n_{ac}}{n_a}$ | $\frac{89}{107} = 0.8318$ |
| Lift | $\left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c}{n}\right)$ | $\left(\frac{89}{107}\right) / \left(\frac{143}{198}\right) = 1.1517$ |
| Leverage | $\frac{n_{ac}}{n} - \frac{n_a}{n} \times \frac{n_c}{n}$ | $\frac{89}{198} - \frac{107}{198} \times \frac{143}{198} = 0.0592$ |
| Importance | $\ln\left[\left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c - n_{ac}}{n - n_a}\right)\right]$ | $\ln\left[\left(\frac{89}{107}\right) / \left(\frac{143 - 89}{198 - 107}\right)\right] = 0.3377$ |
| Conviction | $\frac{n_a \times (n - n_c)}{n \times (n_a - n_{ac})}$ | $\frac{107 \times (198 - 143)}{198 \times (107 - 89)} = 1.6512$ |
| Surprise | $\left(\frac{n_{ac}}{n} - \frac{n_a - n_{ac}}{n}\right) / \left(\frac{n_c}{n}\right)$ | $\left(\frac{89}{198} - \frac{107 - 89}{198}\right) / \left(\frac{143}{198}\right) = 0.4965$ |

In some circumstances, we cannot compute the measure. Thus we use the "error" value "-99.99"

Note: See the following references about these measures

- B. Vaillant, P. Meyer, E. Prudhomme, S. Lallich, P. Lenca, S. Bigaret, « Mesurer l'intérêt des règles d'association », RNTI-E-5, pages 421 - 426, 2006 (in French).
- P. Tan, V. Kumar, J. Srivastava, « Selecting the right interestingness measure for association patterns », Proceedings of 8th ACM SIGKDD, pages 32 – 41, 2002.

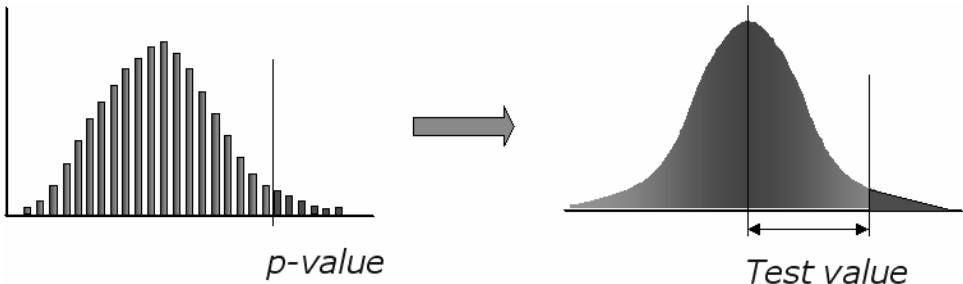
3 Measures based on the "test value" principle ("valeur test" [VT] in French)

The text below is a part of a paper which was published in the Proceedings of the conference EGC-2006 (A. Morineau et R. Rakotomalala, « Critère VT-100 de sélection des règles d'association », Proceedings of the EGC-2006, pages 581 à 592, Lille, 2006 [in French]).

The idea of the test value criterion is to compute the interestingness of a rule from a statistical hypothesis testing scheme. We assume independence between the antecedent and consequent under the null hypothesis. The idea is not to test the absence or the presence of a real link between A and C, but rather to what extent we deviate from the reference situation described by the null hypothesis. In this context, we often compute the p-value of the test.

The problem of the p-value is that it quickly takes very low values, especially when we work on large dataset. It becomes difficult to read and is not really interpretable. To remedy this problem, the concept of value-test was proposed (Morineau, 1984).

The "test value" is the number of standard deviations from the standard Gaussian distribution than we need to cover the computed p-value. Its interpretation is easy: it expresses, in terms of standard deviations, the distance from the reference situation characterized by the null hypothesis.



Let's take a numerical example. If $p = 0.0025$ is the p-value, the test value will be $v = \Phi^{-1}(1 - p) = \Phi^{-1}(1 - 0.0025) = 2.8070$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard Gaussian CDF.

The exact test value (VT FULL) displayed by the A PRIORI MR component corresponds to this framework. Different values according to the null hypothesis and the sampling scheme are supplied.

Here are the results supplied by Tanagra on our data set.

| N° | n | n[A] | n[C] | n[A^C] | VT-Hyp 100 | VT-Hyp Full | VT-Hyp MC | z (Hyp) | VT-Bin 100 (contre-ex.) | VT-Bin Full (contre-ex.) | VT-Bin MC (contre-ex.) | z (contre ex.) | VT-conf 100 | VT-conf Full | VT-conf MC | z (conf) |
|----|-----|------|------|--------|------------|-------------|-----------|---------|-------------------------|--------------------------|------------------------|----------------|-------------|--------------|------------|----------|
| 1 | 198 | 79 | 143 | 68 | 2.275 | 3.478 | 1.768 | 2.281 | 1.657 | 2.542 | 1.227 | 1.920 | 2.216 | 3.069 | 1.764 | 1.954 |
| 2 | 198 | 93 | 143 | 80 | 2.645 | 3.996 | 2.632 | 2.662 | 1.851 | 2.800 | 1.875 | 2.073 | 2.376 | 3.295 | 2.468 | 2.112 |
| 3 | 198 | 74 | 159 | 69 | 2.229 | 3.536 | 2.807 | 2.242 | 1.742 | 2.759 | 2.091 | 2.043 | 2.344 | 3.306 | 2.746 | 1.989 |
| 4 | 198 | 107 | 143 | 89 | 2.403 | 3.586 | 3.057 | 2.416 | 1.563 | 2.348 | 2.237 | 1.798 | 2.021 | 2.763 | 2.842 | 1.798 |
| 5 | 198 | 89 | 143 | 74 | 1.944 | 2.982 | 1.710 | 1.970 | 1.352 | 2.081 | 1.151 | 1.637 | 1.842 | 2.532 | 1.621 | 1.635 |
| 6 | 198 | 89 | 132 | 68 | 1.620 | 2.493 | 2.891 | 1.645 | 1.081 | 1.671 | 1.902 | 1.367 | 1.551 | 2.113 | 2.477 | 1.385 |
| 7 | 198 | 74 | 171 | 71 | 1.802 | 2.997 | 3.614 | 1.847 | 1.436 | 2.390 | 2.950 | 1.856 | 2.079 | 2.957 | 2.026 | 1.707 |
| 8 | 198 | 72 | 171 | 69 | 1.728 | 2.889 | 1.516 | 1.774 | 1.381 | 2.312 | 1.209 | 1.817 | 2.027 | 2.881 | 1.976 | 1.664 |

In order to determine the statistical interest of a rule, we could compare the VT-FULL to the critical value of the test, e.g. 1.65 for a 5% significance level. But we note that it is not really useful in practice. The VT criterion takes a very high value when the size of the database increases. This leads us to propose a normalized test value (VT-100) which we will describe in the next section (Section 4).

Now, we outline the null hypothesis and the sampling scheme of each criterion supplied by Tanagra.

3.1 Hypergeometric sampling scheme (VT-HYP FULL)

In this configuration, the random variable N_{ac} corresponds to the number of examples covered by the rule. The p-value is computed as follows:

$$p = P(N_{ac} \geq n_{ac}) = \sum_{x=n_{ac}}^{\min(n_a, n_c)} \frac{C_{n_c}^x \times C_{n-n_c}^{n_a-x}}{C_n^{n_a}}$$

For the rule n°4 above, we obtain:

$$p = P(N_{ac} \geq 89) = \sum_{x=89}^{\min(107, 143)} \frac{C_{143}^x \times C_{55}^{107-x}}{C_{198}^{107}} = 0.00012195 + \dots = 0.00016788$$

Then the test value is

$$v = \Phi^{-1}(1 - 0.00016788) = 3.586$$

3.2 Counter-examples statistic – Binomial sampling scheme (VT-BIN FULL – Contre-ex.)

$N_{a\bar{c}}$ corresponds to the counter-examples of the rule i.e. the examples covered by the antecedent but not the consequent. We use a binomial scheme. Under the null hypothesis, the probability of a counter-example is

$$\pi = \frac{n_a \times (n - n_c)}{n^2}$$

The p-value is computed as follows

$$p = P(N_{a\bar{c}} \leq n_{a\bar{c}}) = \sum_{x=0}^{n_{a\bar{c}}} C_n^x \pi^x (1 - \pi)^{n-x}$$

For the rule n°4,

$$p = P(N_{a\bar{c}} \leq n_{a\bar{c}}) = \dots + C_{198}^{18} 0.15^{18} (1 - 0.15)^{198-18} = 0.009445$$

and

$$v = \Phi^{-1}(1 - 0.009445) = 2.348$$

3.3 Confidence statistic – Binomial scheme (VT-BIN CONF.)

We use now the binomial sampling scheme for the N_{ac} statistic. Under the null hypothesis, we have

$$\pi = \frac{n_c}{n}$$

We obtain the p-value

$$p = P(N_{ac} > n_{ac}) = \sum_{x=n_{ac}+1}^{n_a} C_{n_a}^x \pi^x (1-\pi)^{n_a-x}$$

For the rule n°4,

$$p = P(N_{ac} > n_{ac}) = C_{107}^{90} 0.72^{90} (1-0.72)^{107-90} + \dots = 0.002860$$

The test value is

$$v = \Phi^{-1}(1 - 0.002860) = 2.763$$

4 The normalized test value – La VT-100

The main drawback of the test value is to take a very high value when the size of the database increases. We can use the test value to rank the rules, but we cannot use it to detect the “relevant” rules.

In order to alleviate this undesirable behavior, the normalized test value is proposed. The size of the examples in the data table above is arbitrarily coerced to 100.

"100" is of course a very experimental value. The motivation and the empirical justification of this value can be found into various references, especially (Rakotomalala and Morineau, 2008; “The TVpercent principle for the counterexamples statistic”, in *Statistical Implicative Analysis, Studies in Computational Intelligence Series*, 127, 449-462, 2008 -- <http://www.springerlink.com/content/g245317206950529/>).

The VT-100 measure can be computed in two ways:

(1) We use a repeated random sampling approach (MC: Monte-Carlo). At each attempt, we draw a sample of 100 examples. Then we compute the mean of the test value. In the A PRIORI MR component, REPETITION allows to set the number of replications. This approach can be very expensive if we set a high number of replications. It can be unstable if we set a low number of replications.

(2) The other way is to coerce the values into the table to 100. Because we are able to obtain fractional values, which is not suitable for the chosen sampling scheme, the final test value is obtained by averaging the test value of the discrete neighbors of the studied configuration.

For the rule n°4, the data table is

| Tableau | A | Non (A) | Total |
|----------------|-------|---------|--------|
| C | 44.95 | 27.27 | 72.22 |
| Non (C) | 9.09 | 18.69 | 27.78 |
| Total | 54.04 | 45.96 | 100.00 |

We obtain the following results according the different approaches

| | VT 100 | VT MC | VT FULL |
|---------|--------|-------|---------|
| VT-HYP | 2.403 | 2.426 | 3.586 |
| VT-BIN | 1.563 | 1.557 | 2.348 |
| VT-CONF | 2.021 | 2.014 | 2.763 |

We note two main results:

- The obtained measures (VT-100 and VT MC) are rather similar.
- Because the sample size is artificially divided by 2 (from $n = 198$, we set $n = 100$), in comparison to the VT-FULL, the normalized test value is approximately divided by $\sqrt{2}$

5 Z-value based on the Gaussian approximation

When the size of the database is sufficiently high, we can use the Gaussian approximation of the Binomial and the Hypergeometric distributions. We can then compute directly the z-value which is another approximation of the test value.

According to the sampling scheme, we obtain:

- Z-HYP :
$$z_{hyp} = \frac{n_{ac} - n_a n_c / n}{\sqrt{\frac{(n_a (n - n_a) / n)(n_c (n - n_c) / n)}{n - 1}}}$$
- Z-BIN (counter-examples) :
$$z_{contre-ex} = \frac{n_{ac} - n_a n_c / n}{\sqrt{n(n_a n_c / n^2)(1 - n_a n_c / n^2)}}$$
- Z-CONF :
$$z_{conf} = \frac{n_{ac} - n_a n_c / n}{\sqrt{n_a \frac{n_c}{n} (1 - \frac{n_c}{n})}}$$

This measure can be computed on the whole dataset or on the “coerced to 100” dataset. In this configuration, we have an approximation of the VT-100 measure.

| | Z | VT 100 |
|---------|-------|--------|
| VT-HYP | 2.639 | 2.403 |
| VT-BIN | 1.658 | 1.563 |
| VT-CONF | 1.798 | 2.021 |

The approximation is maybe better if we use the correction of continuity (Yates). It is not implemented in the version 1.4.30 of Tanagra.

6 Partitioning the dataset into train and test set

According to the supervised learning framework, we can subdivide the dataset in a learning set which is used to compute the rules, and a test set, which is used to assess the rules.

The LEARNING SET RATIO option allows setting the proportion of the dataset used in the learning phase. If we set this parameter to 1, it means that the whole dataset is used in the learning phase.

7 Conclusion

The A PRIORI MR and the SPV ASSOC RULE components are experimental tools for the evaluation of the rules extracted by the association rule induction algorithm. They allow to evaluate the rules using measures based on the test value principle.