

Ce document recense les mesures d'évaluation des règles d'association proposées par le composant A PRIORI MR. Elles résultent d'études relatées dans une série de publications de A. Morineau et R. Rakotomalala (essentiellement en 2006, voir <http://eric.univ-lyon2.fr/~ricco/publications.html>).

Une mesure sert à caractériser la pertinence d'une règle. Elle permet de les classer. Elle devrait aussi permettre de discerner celles qui sont « significativement intéressantes » de celles qui ne le sont pas. Ce dernier point reste totalement prospectif. Il n'y a pas de solutions réellement satisfaisantes à ce jour.

1 Tableau de travail

Tout d'abord, voici quelques indications sur les notations utilisées. Une règle est composée d'un antécédent et d'un conséquent, lesquels sont formés d'une liste d'items. Nous pouvons résumer les effectifs des individus couverts par une règle à l'aide du tableau suivant :

	Antécédent	Non (Antécédent)	Effectif
Conséquent	n_{ac}		n_c
Non (Conséquent)			
Effectif	n_a		n

Avec :

- n est le nombre d'observations dans la base complète ;
- n_a est le nombre d'observations couvertes par l'antécédent de la règle ;
- n_c est le nombre d'observations couvertes par le conséquent de la règle ;
- n_{ac} est le nombre d'observations couvertes par la règle c.-à-d. à la fois par l'antécédent et le conséquent.

Dans la copie d'écran ci-dessous, nous reprenons les résultats de Tanagra sur un exemple.

N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise
1	"habit=locataire" - "csp=cadre_moyen" - "port_action=oui"	"accord=oui"	198	79	143	68	0.3434	0.8608	1.1918	0.0553	0.3117	1.9949	0.3986
2	"habit=locataire" - "port_action=oui"	"accord=oui"	198	93	143	80	0.4040	0.8602	1.1911	0.0648	0.3603	1.9872	0.4685
3	"Age=ancien"	"habit=locataire" - "csp=cadre_moyen"	198	74	159	69	0.3485	0.9324	1.1611	0.0484	0.2505	2.9152	0.4025
4	"port_action=oui" "csp=cadre_moyen"	"accord=oui"	198	107	143	89	0.4495	0.8318	1.1517	0.0592	0.3377	1.6512	0.4965

Pour la règle n° 4 : « SI port action = oui ALORS accord = oui », « port action = oui » constitue l'antécédent, « accord = oui » le conséquent. Nous retrouvons les données suivantes :

	Antécédent	Non (Antécédent)	Effectif
Conséquent	89		143
Non (Conséquent)			
Effectif	107		198

Les indicateurs affichés par Tanagra sont calculés à partir de ces données.

2 Mesures « classiques » d'évaluation des règles

Nous qualifions de « classiques » ces mesures pour les différencier des indicateurs basés sur la notion de valeur test que nous présenterons plus loin.

Mesure	Formule	Valeur
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	$\ln\left[\frac{\left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c - n_{ac}}{n - n_a}\right)}{\left(\frac{89}{107}\right) / \left(\frac{143 - 89}{198 - 107}\right)}\right]$	$\ln\left[\frac{\left(\frac{89}{107}\right) / \left(\frac{143 - 89}{198 - 107}\right)}{\left(\frac{89}{107}\right) / \left(\frac{143 - 89}{198 - 107}\right)}\right] = 0.3377$
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	$\left(\frac{n_{ac}}{n} - \frac{n_a - n_{ac}}{n}\right) / \left(\frac{n_c}{n}\right)$	$\left(\frac{89}{198} - \frac{107 - 89}{198}\right) / \left(\frac{143}{198}\right) = 0.4965$

Dans certaines situations, il n'est pas possible de calculer l'indicateur. Tanagra renvoie alors le code d'erreur -99.99.

Remarque : Pour un tour d'horizon étendu des mesures d'évaluation des règles d'association, voir

- 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.
- P. Tan, V. Kumar, J. Srivastava, « Selecting the right interestingness measure for association patterns », Proceedings of 8th ACM SIGKDD, pages 32 – 41, 2002.

3 Mesures exactes basées sur la notion de valeur-test

Ce qui suit reprend en partie un texte qui a été présenté à la conférence EGC-2006 (A. Morineau et R. Rakotomalala, « Critère VT-100 de sélection des règles d'association », Actes de EGC-2006, pages 581 à 592, Lille, 2006).

L'idée de la valeur-test est de calculer l'intérêt d'une règle à partir d'un schéma de test statistique. On postule l'indépendance entre l'antécédent et le conséquent sous l'hypothèse nulle.

Lors de sa mise en oeuvre, le statisticien calcule la probabilité critique (p-value). Il l'utilise de deux manières : il la confronte avec un risque d'erreur, défini a priori lors de la mise en place du test (ex. 5%), pour savoir si l'écart par rapport à l'hypothèse nulle est statistiquement significatif ; il s'en sert pour caractériser l'éloignement par rapport à la situation de référence que constitue l'hypothèse nulle. Plus faible sera la p-value, plus fort sera le rejet de l'hypothèse nulle.

Le problème de la p-value est qu'elle prend rapidement des valeurs très faibles, surtout lorsque l'on travaille sur de grands effectifs. Elle devient peu lisible et n'est pas vraiment interprétable. Pour remédier à cet écueil, la notion de valeur-test a été proposée (Morineau, 1984).

La « valeur-test » correspond au nombre d'écarts type de la loi normale centrée réduite qu'il faut dépasser pour couvrir la p-value calculée. Son interprétation facile est son principal intérêt : on exprime en termes d'écarts type l'éloignement par rapport à la situation de référence caractérisée par l'hypothèse nulle.

Exemple : Prenons un exemple simple pour fixer les idées : si $p = 0.0025$ est la p-value, la valeur test associée sera $v = \Phi^{-1}(1 - p) = \Phi^{-1}(1 - 0.0025) = 2.8070$, où $\Phi^{-1}(\cdot)$ est l'inverse de la fonction de répartition de la loi normale centrée et réduite.

Les valeurs test exactes (FULL) proposées par le composant A PRIORI MR correspondent donc à des valeurs déduites des probabilités critiques des tests d'hypothèses mis en place pour caractériser les règles. Elles diffèrent par la statistique du test, l'écriture de l'hypothèse nulle et le schéma d'échantillonnage.

Voici le tableau de résultats fourni par Tanagra sur notre jeu de données.

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

Pour déterminer l'intérêt statistique d'une règle, on pourrait comparer la valeur test avec les seuils critiques usuels que l'on utilise lorsque la statistique de test distribuée selon une loi normale centrée et réduite. Pour un test à 5%, le seuil est approximativement égal à 1.65. On se rend compte à l'usage que ce n'est pas tenable. La valeur test, tout comme la p-value, a tendance à prendre des valeurs extrêmes sur les grands effectifs. La démarche n'est pas appropriée pour éliminer les règles qui seraient inintéressantes. Nous reviendrons sur ce sujet plus loin lorsque nous présenterons la VT-100.

3.1 Tirage hypergéométrique (VT-HYP FULL)

Dans cette configuration, nous travaillons à marges fixées c.-à-d. les effectifs marginaux ne sont pas aléatoires. On considère que sous l'hypothèse nulle, la variable aléatoire N_{ac} , qui représente le nombre d'individus couverts par la règle, suit une loi hypergéométrique. La probabilité critique du test est calculée de la manière suivante :

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

Prenons la règle n°4 ci-dessus. La probabilité critique du test s'écrit :

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

Nous en déduisons la valeur test

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

3.2 Statistique du contre-exemple - Tirage binomial (VT-BIN FULL – Contre-ex.)

Nous modélisons le nombre de contre-exemples N_{ac} à l'aide de la distribution binomiale. Sous l'hypothèse nulle, la probabilité d'occurrence des contre-exemples est égale à

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

La probabilité critique du test s'écrit

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

Pour la règle n°4, nous obtenons

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

Nous obtenons ainsi

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

3.3 Statistique de confiance – Tirage binomial (VT-BIN CONF.)

Nous modélisons le nombre d'individus couverts par la règle N_{ac} à l'aide d'un schéma binomial. Nous comparons cette fois les profils colonnes dans notre tableau de comptage. Ainsi, la caractérisation de la situation de référence est différente, sous l'hypothèse nulle la probabilité d'occurrence des exemples s'écrit

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

La probabilité critique est calculée de la manière suivante

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

Pour la règle n°4, nous avons

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

Et la valeur test

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

4 Valeur test ramenée à 100 – La VT-100

La valeur-test est un indicateur statistique. De fait, il présente l'avantage d'être croissant avec les effectifs. Toutes choses égales par ailleurs (c.-à-d. les proportions étant les mêmes dans notre tableau de calcul), le résultat sera d'autant plus crédible que la base de données est plus grande.

Mais c'est aussi son principal inconvénient. Très rapidement, l'indicateur prend des valeurs que l'on ne peut plus confronter aux seuils usuels. Il permet bien de comparer les règles entre elles, de les classer selon leur pertinence. En revanche, nous ne pouvons pas nous en servir pour éliminer les règles qui ne seraient pas significatives.

Pire, dans certains cas, lorsque les effectifs sont très élevés, la p-value n'est plus calculable à l'aide des bibliothèques numériques disponibles¹. L'obtention de la valeur test n'est plus possible.

Pour palier cet inconvénient, le principe de la VT-100 a été avancé. L'idée est de ramener arbitrairement la taille d'échantillon à 100 observations². Nous pouvons le faire de 2 manières.

(1) Nous utilisons une procédure de Monte-Carlo c.-à-d. un tirage aléatoire avec remise de 100 observations que nous répétons. A chaque tirage, nous calculons la valeur test. La valeur test Monte-Carlo (VT MC) correspond à la moyenne des valeurs obtenues. Dans A PRIORI MR, le paramètre REPETITION permet de définir le nombre de tirages à réaliser lors de la procédure. Attention, si le nombre de répétition est élevé, les calculs peuvent être très longs ; s'il est faible, les résultats peuvent être très instables.

(2) L'autre idée est de ramener arbitrairement les effectifs à $n = 100$ (VT 100). L'ennui est que nous pouvons obtenir des valeurs fractionnaires dans certaines cases du tableau. C'est le cas pour la règle n°4.

¹ J'ai utilisé intensivement le package TURBOPOWER SYSTOOLS lors des expérimentations. Je n'ai pas trouvé de bibliothèque (libre) plus performante en Pascal → <http://sourceforge.net/projects/tpsystools/>

² Pourquoi 100, pourquoi pas 100... vaste discussion que nous n'aborderons par ici. Le plus simple est de se rapporter aux références indiquées dans le texte. L'objectif est de produire un indicateur que l'on peut confronter aux seuils usuels de la statistique (1.65 pour un test à 5%, 1.96 pour un test à 2.5%, 2.33 à 1%, etc.). Ceci étant, je reconnais qu'il y a une énorme part d'empirisme là dedans. Mais bon, quand on fixe un risque de première espèce à 5%, n'est ce pas non plus empirique ? Pourquoi 5%, pourquoi pas 25% ?

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

Dans ce cas, la stratégie consiste à calculer la valeur test par interpolation, à partir de la moyenne pondérée des configurations de valeurs entières situées dans son voisinage. C'est un peu une usine à gaz. Mais les expérimentations montrent que cette procédure permet de se rapprocher au mieux des résultats de la procédure de Monte-Carlo avec un nombre élevé de répétitions.

Pour la règle n°4, selon la modélisation adoptée, nous obtenons

	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

Nous remarquons 2 points importants :

- Les résultats sont assez proches. Ils le seront d'autant plus que nous augmenterons le nombre de répétitions pour la procédure de Monte-Carlo.
- Les effectifs ayant été artificiellement divisés par 2 (de n = 198, on passe à n = 100), par rapport aux VT-FULL, les valeurs tests sont mécaniquement divisés par $\sqrt{2}$ (approximativement).

5 Mesures approchées basées sur la notion de valeur-test

Lorsque les effectifs sont suffisamment élevés, il est possible de passer par l'approximation normale des lois hypergéométriques et binomiales. Nous pouvons ainsi approximer directement la valeur test à partir des statistiques centrées réduites des tests (Z), sans qu'il soit nécessaire de passer par le calcul de la p-value. Le calcul est plus rapide et mieux sécurisé.

Selon la configuration utilisée, nous obtenons Z de la manière suivante :

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

Cet indicateur peut être obtenu à partir du tableau de calcul initial, nous pouvons aussi, et c'est ce qui est fait dans Tanagra, le calculer à partir du tableau des effectifs ramenés à 100 pour approximer la VT-100.

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

L'approximation pourrait être meilleure avec la correction de continuité (correction de Yates) mais nous ne l'introduisons pas dans la version 1.4.30 de Tanagra.

6 Apprentissage et test

A PRIORI MR intègre une option très intéressante. Il est possible de scinder aléatoirement en 2 parties le fichier de données : la première correspond à l'ensemble d'apprentissage, il sert à produire les règles ; la seconde correspond à l'ensemble test, il sert à évaluer les règles.

LEARNING SET RATIO permet de régler les proportions respectives. Lorsqu'il est égal à 1, cela indique que nous réservons la totalité des données pour l'apprentissage.

L'objectif est de détecter les situations de sur apprentissage. Il apparaît à l'usage que cette option n'est pas très déterminante. En effet, le système de construction des règles ne cherche pas à optimiser explicitement un critère lors de la phase d'apprentissage (ni la valeur test, ni les autres). Certains critères (support, confiance, lift aussi dans A PRIORI MR) n'agissent que pour filtrer les règles ou limiter l'espace de recherche. De fait, la qualité de la règle est rarement sur-évaluée sur les données en apprentissage.

7 Conclusion

A PRIORI MR est un composant expérimental. Nous nous en sommes surtout servis pour analyser le comportement des différentes mesures d'évaluations des règles d'association, notamment pour positionner la mesure « valeur test » par rapport aux indicateurs reconnus de la littérature.

J'avoue l'avoir un peu oublié. Je me suis replongé dedans très récemment suite à un commentaire d'un internaute qui se demandait à quoi correspondait toutes ces colonnes de chiffres qui apparaissent dans le tableau de résultats.