Critère VT100 de sélection des règles d'association

Alain Morineau*, Ricco Rakotomalala**

*MODULAD, Paris
alain.morineau@modulad.fr
http://www.modulad.fr
**Laboratoire ERIC – Université Lyon 2
ricco.rakotomalala@univ-lyon2.fr
http://eric.univ-lyon2.fr/~ricco

Résumé. L'extraction de règles d'association génère souvent un grand nombre de règles. Pour les classer et les valider, de nombreuses mesures statistiques ont été proposées ; elles permettent de mettre en avant telles ou telles caractéristiques des règles extraites. Elles ont pour point commun d'être fonction croissante du nombre de transactions et aboutissent bien souvent à l'acceptation de toutes les règles lorsque la base de données est de grande taille. Dans cet article, nous proposons une mesure inspirée de la notion de valeur-test. Elle présente comme principale caractéristique d'être insensible à la taille de la base, évitant ainsi l'écueil des règles fallacieusement significatives. Elle permet également de mettre sur un même pied, et donc de les comparer, des règles qui auront été extraites de bases de données différentes. Elle permet enfin de gérer différents seuils de signification des règles. Le comportement de la mesure est détaillé sur un exemple.

1 Introduction

1.1 Les valeurs-tests

Pour faire un test de l'hypothèse nulle H_0, le statisticien calcule une « *probabilité critique* » (ou *p-value*). C'est la probabilité, calculée sous H_0, d'un événement au moins aussi extrême que l'événement observé. De façon intuitive, on comprend que cette probabilité est d'autant plus faible qu'on est loin de l'hypothèse nulle. Si l'événement observé est très improbable sous l'hypothèse nulle, on jugera que les observations sont vraisemblablement régies par un mécanisme *non nul*. Il est donc tentant d'utiliser cette valeur numérique pour évaluer l'écart entre ce qu'on a observé et la situation « sans intérêt » correspondant à ce qu'on aurait observé sous H_0. Dans ce contexte, plus l'évaluation de l'écart est forte (plus la *probabilité critique* est faible), plus ce qu'on a observé est *intéressant* (Gras *et al.*, 2002 ; Lerman et Azé, 2003 ; Lallich et Teytaud, 2004). Dans la pratique, on se rend compte que la p-value est difficile à manipuler ; elle peut atteindre des valeurs très faibles, très peu lisibles ; pire, dans certains cas, elle est inutilisable car on se heurte aux limites de l'approximation

des fonctions de répartition des lois de probabilités d'usage courant à l'aide des bibliothèques de calcul couramment utilisées par la communauté scientifique.

Pour avoir une mesure mieux adaptée (en particulier, qui soit croissante si l'écart est croissant), on remplace cette probabilité critique par le nombre d'écarts types de la loi normale centrée et réduite qu'il faut dépasser pour couvrir cette probabilité. On a appelé *valeur-test* cette mesure exprimée en nombre d'écarts types (Morineau, 1984 ; Lebart *et al.*, 1995) et on l'a exploitée de façon intensive dans le logiciel *SPAD* pour la partie dévolue au Data mining.

1.2 Un critère numérique de classement et de sélection

La notion de *critère* invoquée ici est particulière. On utilise en effet les outils mis en œuvre dans les tests d'hypothèses mais il ne s'agit en aucun cas de faire des tests au sens de la théorie classique des tests statistiques. L'objectif n'est pas, par exemple, de tester l'indépendance entre A et C puisqu'on tiendra pour vrai que A et C sont liés. La mécanique des tests est en quelque sorte dévoyée pour servir comme *outil d'évaluation* et non comme outil de prise de décision dans l'incertain.

On se place donc délibérément hors du cadre de la théorie de la décision statistique, et par conséquent hors d'atteinte des critiques qui s'y rapporteraient. Cette remarque doit écarter à l'avance les objections que ne manquerait pas de faire à juste titre le statisticien qui lirait ces lignes sans ce préalable.

1.3 Les comparaisons multiples

Comme on le voit dans tout contexte de *Data mining* et en particulier dans le logiciel *SPAD*, les valeurs-tests sont calculées pour évaluer, sur le même jeu de données, des dizaines ou des centaines d'écarts entre moyennes ou entre pourcentages, ou pour évaluer autant de corrélations entre variables mesurées sur les mêmes individus.

Dans le cadre de l'application classique des tests statistiques (on accepte de rejeter à tort l'hypothèse nulle avec une probabilité fixée α), une telle situation nécessiterait de corriger le seuil confiance de chaque test pour assurer un risque global fixé à α (correction de *Bonferroni* ou toute autre correction proposée dans la littérature).

S'il s'agit d'évaluer dans quelle mesure les observations supportent l'idée qu'on est *éloigné d'une hypothèse nulle*, il n'y a pas de seuil de confiance en jeu : on évalue simplement des écarts relatifs entre les valeurs-tests. Il y a effectivement des comparaisons multiples mais la question de la correction pour comparaisons multiples ne se pose pas dans ce contexte.

1.4 Les inconvénients de la valeur-test

Evaluer la force ou l'intérêt des liaisons et des écarts est un problème qu'on rencontre dans le contexte du Data mining où le nombre des observations est toujours « très grand », bien au-delà des tailles d'échantillons rencontrés dans une situation de test d'hypothèse. Les *probabilités critiques* seront par conséquent généralement très faibles, et d'ailleurs souvent difficiles à calculer pour cette raison. En d'autres termes, les *valeurs tests* seront très grandes et toujours, à phénomène égal, fonctions croissantes du nombre d'observations.

Illustrons cette remarque sur un exemple où l'on compare deux pourcentages en utilisant par exemple la loi hypergéométrique (on place les effectifs dans un tableau 2x2 dont les marges sont supposées fixées).

Considérons un groupe particulier d'individus représentant 56% des observations totales. Considérons un caractère qui est présent dans 23% de l'ensemble de ces observations, mais dans 28% du groupe particulier (le groupe noté A). Le même écart entre 23% et 28% correspond à une valeur test qui dépend évidemment du nombre des observations sur lesquelles il est calculé. Choisissons quelques cas correspondant à ces mêmes pourcentages pour des tailles croissantes, de n=100 à n=4000. Les données sont présentées dans les tableaux 2x2 de la *Figure 1*.

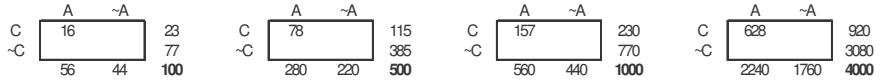

FIG. 1 -- *Effectifs croissant pour un même phénomène*

Les ordres de grandeur des valeurs-tests (loi hypergéométrique) sont :

Taille	Valeur-test
100	1.3
500	2.8
1000	4.2
4000	8.5

FIG. 2 -- *Exemple de croissance des valeurs-tests*

Ce phénomène est bien connu des statisticiens : si la taille de l'échantillon croît suffisamment, le moindre écart à toute hypothèse nulle finit par devenir « significatif ».

Dans ces conditions, non seulement les valeurs-tests ne mesurent pas la « significativité » d'un écart, mais de plus elles ne peuvent pas être comparées si elles sont calculées sur des nombres différents d'observations.

2 La Valeur-test normalisée VT100 pour les règles d'association

On ne rappellera pas ici les nombreux travaux déjà publiés qui s'attaquent au problème de détection des règles d'association *« intéressantes »* parmi la liste toujours longue des règles fournies par les algorithmes classiques, comme l'algorithme *A PRIORI* (Agrawal et Srikant, 1994) ou ses divers avatars. On pourra utilement consulter certains articles de synthèse (Lenca *et al.*, 2003 ; Tan *et al.*, 2002).

Les remarques faites ci-dessus nous conduisent à proposer une sorte de *valeur-test normalisée*, jouant un rôle analogue à celui de la valeur-test habituelle, mais rendue *indépendante du nombre d'observations* sur lesquelles elle est calculée. A ce titre, son rôle essentiel sera de ranger les règles d'association selon leur ordre d'intérêt et accessoirement de suggé-

rer un seuil en deçà duquel les règles n'auront vraisemblablement plus d'intérêt (quelque soit la taille des données).

Le principe en est simple : on décide de se ramener artificiellement au cas du nombre d'observations n=100. Cette valeur – a priori arbitraire - étant à rapprocher de la taille *raisonnable* des échantillons utilisés quand la théorie des tests s'est historiquement développée. Pour ce faire, on imagine le mécanisme suivant.

Considérons une règle d'association particulière construite sur les n observations disponibles (n étant grand comparé à 100). Il lui correspond un tableau 2x2 où la somme des fréquences vaut n. Comme on l'a introduit plus haut, on mesure l'intérêt de la règle en s'appuyant sur l'écart à l'indépendance mesuré à partir de la loi hypergéométrique (on verra plus loin que le raisonnement sera analogue pour tout autre critère).

Dans le contexte hypergéométrique, les marges du tableau 2x2 sont fixées. On imagine donc les données comme étant des 0 et des 1 répartis dans deux colonnes à n composantes, de façon aléatoire, le nombre de 1 étant fixé dans chaque colonne. Un échantillon de taille 100 extrait de ces données correspond à un tirage au hasard de 100 lignes dans ce tableau à 2 colonnes de longueur n.

On répète un grand nombre de fois le tirage d'échantillons de taille 100. Pour chaque tirage, on construit le tableau 2x2 correspondant et on calcule la probabilité critique hypergéométrique. Finalement on calcule la probabilité critique moyenne (la *p-value* moyenne) que l'on obtient à partir de ces échantillons. Appelons VT_{100} la valeur-test associée à cette probabilité critique moyenne. Cette moyenne VT_{100} sera le critère utilisé pour caractériser l'intérêt de la règle.

Ce mécanisme de calcul du critère VT_{100} tend à placer l'utilisateur dans le contexte classique d'un échantillon aléatoire de taille n=100 porteur de l'association à évaluer. On serait précisément dans ce contexte classique si la population réelle était l'ensemble des observations et si on ne tirait qu'un seul échantillon. En tirant de nombreux échantillons, on stabilise la valeur de la *p value* tout en conservant la taille 100 du support de calcul. Dans ce contexte, on ne s'interdira pas de penser ainsi : « *si je trouve une VT_{100} très supérieure à 1,645 (car 1,645 écart type est le seuil des valeurs tests d'un test unilatéral à 5%), je suis vraisemblablement en présence d'une règle très intéressante* ».

Ainsi le critère VT_{100} permettra de ranger par ordre d'intérêt les règles calculées sur une même base de données. Il présente aussi l'avantage de permettre la comparaison de règles calculées sur des bases de données différentes, par exemple sur des bases de données de tailles différentes extraites à des dates successives.

3 Calcul pratique des VT100

Le mécanisme de tirage répété des échantillons de taille 100 est utile pour présenter le critère. Pour réaliser les calculs, on utilisera une procédure d'approximation rapide et suffisamment efficace : rapide, car elle ne nécessite pas la répétition des tirages, et donc des accès à la base ; efficace, car elle fournit des valeurs proches de celles du tirage répété et permet de classer les règles dans le même ordre. On suivra facilement cette procédure de calcul sur un exemple présenté en *Figure 3*. Considérons une règle A =>C définie dans le tableau 2x2 ci-dessous sur 2000 observations. On calcule le tableau correspondant des effectifs (décimaux) ramenés à un total égal à 100.

A=>C	A	~A	
C	226		568
~C			
	346		**2000**

A=>C	A	~A	
C	11,30		28,40
~C			
	17,30		**100**

FIG. 3 -- *Les effectifs sont décimaux*

Il s'agit d'approcher par interpolation ce que donnerait une loi hypergéométrique appliquée à ce tableau d'effectifs décimaux. Chacun des trois effectifs décimaux écrit dans le tableau est compris entre deux entiers proches (par exemple 11,30 est compris entre 11 et 12). Ceci conduit à imaginer d'approcher le résultat cherché comme barycentre des résultats hypergéométriques calculés sur les 8 tableaux obtenus en combinant les effectifs entiers les plus proches des 3 valeurs décimales, les coefficients barycentriques étant les écarts aux entiers.

L'approximation peut paraître naïve mais elle est certainement suffisante pour l'objectif à atteindre qui est de comparer l'intérêt des règles. Rien n'empêche d'ailleurs de lui substituer à volonté un calcul plus raffiné pourvu que le temps de calcul reste raisonnable (noter que, la taille 100 étant fixée, la loi hypergéométrique à calculer n'a plus que 2 paramètres entiers libres, ce qui suggère l'alternative d'une tabulation à double entrée, stockée en mémoire pour éviter tout calcul hypergéométrique dans les applications).

4 Un même principe pour différentes mesures d'intérêt

Lorsqu'on se base sur la loi hypergéométrique pour évaluer l'intérêt d'une règle A=>C (c'est-à-dire pour évaluer l'écart à une situation d'indépendance entre A et C), le critère utilisé est symétrique en A et C et donnera le même résultat pour la règle C=>A. L'intérêt de la règle est mesurée en fait par l'écart entre son *support* et ce que serait ce support en cas d'indépendance de A et de C (si les effectifs respectifs de A et C sont considérés comme des quantités fixées). A ce titre, le critère VT_{100} basé sur la loi hypergéométrique pourrait être considéré comme un versant *statistique* du critère numérique usuel appelé *Lift* dans le cadre des règles d'association.

Il y a bien sûr de nombreux autres points de vue possibles pour apprécier l'intérêt d'une règle. A titre d'exemples, nous allons évoquer deux autres utilisations du critère VT_{100}. Dans tous les cas, il est commode de se représenter les observations concernant une règle A=>C comme 2 colonnes de longueur n, nommées A et C, contenant des 0 et des 1. Le cas d'indépendance déjà évoqué associé à la loi hypergéométrique consiste à répartir de façon aléatoire un nombre n(A) fixé de 1 dans la colonne A et un nombre n(C) fixé de 1 dans la colonne C.

4.1 Critère VT_{100} associé à la confiance d'une règle

Un autre contexte d'indépendance entre A et C est défini par le mécanisme suivant : dans la colonne A , les 1 sont introduits avec la probabilité constante p(A) et, de façon indépendante, les 1 sont introduits dans la colonne C avec la probabilité constante p(C). Dans ces conditions (schéma dit *binomial*), les marges du tableau 2x2 ne sont pas fixées.

Une règle A=>C est d'autant plus intéressante que sa *confiance*, fréquence du conséquent sachant que l'antécédent est réalisé, calculée par n(AetC)/n(A), s'écarte davantage de la

fréquence globale du conséquent n(C)/n. Avec le schéma binomial, et si on estime la probabilité p(C) du conséquent par sa fréquence empirique n(C)/n, la loi du support n(AetC) est la loi binomiale de paramètres p(C) et n(A).

La probabilité critique d'observer, sous l'hypothèse d'indépendance (loi binomiale), une confiance au moins aussi extrême que celle qu'on a observée servira à mesurer l'intérêt de la règle en terme de *confiance*. Transformée en nombre d'écarts types d'une loi normale, cette probabilité critique devient la valeur-test. Ce critère évaluant l'intérêt d'une règle possède la particularité de ne pas être symétrique (il est différent pour la règle C=>A) mais, comme le précédent, il a l'inconvénient d'être sensible aux effectifs.

Selon le principe présenté plus haut, on le normalise en ramenant le paramètre n(A) à la valeur 100. Dans ces conditions, on se trouve en présence d'une loi binomiale dont les paramètres sont connus (taille 100 et probabilité n(C)/n) mais il faudrait calculer la probabilité de dépasser une valeur *non entière* n(AetC)/n(A). L'approximation de cette valeur peut se faire par interpolation comme précédemment : on calcule les probabilités critiques binomiales pour les deux entiers qui encadrent la valeur décimale et on approche la valeur cherchée par leur barycentre en prenant comme coefficients les écarts aux entiers. La transformation de cette probabilité en nombre d'écarts types de la loi normale sera le critère VT_{100} associé à la confiance de la règle.

On signale une propriété satisfaisante de ce critère (partagée bien sûr par d'autres critères). Considérons deux règles ayant le même antécédent A et les conséquents C_1 et C_2. Supposons que ces deux règles aient la même confiance, donc n(Aet C_1) est égal à n(Aet C_2). Si n(C_1) est inférieur à n(C_2), la règle A=> C_1 sera préférée car elle creuse l'écart entre la confiance (qui est la même) et la fréquence du conséquent. Il est clair que le critère VT_{100} calculé à partir de la loi binomiale favorisera bien la règle A=> C_1.

4.2 Critère VT_{100} associé aux contre-exemples

Un contre-exemple de la règle A=>C consiste en la réalisation de A alors que C n'est pas réalisé. L'intérêt pour une règle est d'autant plus grand que le nombre de contre-exemples noté n(Aet~C) est faible. Dans le schéma binomial précédent (répartition des 0 et des 1 dans les colonnes A et C avec des probabilités constantes), l'hypothèse d'indépendance implique une loi simple pour le nombre X de contre-exemples. On peut en effet estimer la probabilité d'un contre-exemple par le produit des probabilités des deux événements indépendants p = {n(A)/n} {n(~C)/n}. Ainsi X suivra une loi binomiale de paramètres connus n et p.

L'intérêt d'une règle mesuré sous l'angle des contre-exemples reposera sur la probabilité, calculée sous l'hypothèse d'indépendance, d'un nombre de contre-exemples au moins aussi extrêmes (c'est-à-dire, plus petit) que le nombre observé. Cette probabilité sera évaluée dans le cadre normé d'un échantillon de taille 100 puis transformée en nombre d'écarts types d'une loi normale pour définir la VT_{100} associée aux contre-exemples. Ramené à 100 observations, le nombre de contre-exemples devient une valeur décimale ; on approchera donc la probabilité critique cherchée par le barycentre des deux probabilités binomiales calculées pour les entiers qui l'encadrent.

Parmi les propriétés de ce critère, on notera qu'il est non symétrique en A et C et qu'il attribue la même valeur à la règle A=>C et à sa *contraposée* qui lui est logiquement équivalente ~C=>~A.

Remarque

Puisque les effectifs utilisés sont supposés ici très élevés, les distinctions entre certaines situations d'indépendance peuvent s'estomper et les distributions de probabilités (hypergéométriques, binomiales, khi-2 …) tendre vers les mêmes lois limites. Ceci peut faire apparaître assez artificielles dans certains cas les distinctions faites ici entre les schémas d'indépendance.

5 Un exemple d'application

5.1 Données et résultats

Les données utilisées sont extraites du fichier « Adult » disponible sur le site « UCI Machine Learning Repository » (Newman et al., 1998). Les données manquantes ont été remplacées par la moyenne pour les attributs continus, par une nouvelle modalité « manquante » pour les attributs discrets. Le tableau analysé ici possède 14 743 lignes (individus) et 12 colonnes (variables qualitatives dont certaines sont les variables quantitatives d'origine recodées en classes). La dernière variable est une variable Oui/Non indiquant si l'individu a un gain moyen supérieur ou non à US$ 50000. On s'intéresse aux règles dont le conséquent est l'attribut Non de cette variable et dont les antécédents peuvent contenir jusqu'à 3 items. Pour la sélection des règles par un algorithme classique « A PRIORI », on s'est fixé comme seuils un support supérieur à 20%, une confiance supérieure à 60% et un lift supérieur à 1,10. L'implémentation correspond au composant « A PRIORI MR » dans le logiciel TANAGRA (Rakotomalala, 2005), le code source est disponible en ligne.

Pour la clarté de l'illustration numérique, on présente une application "supervisée" où les règles doivent conduire à un conséquent unique choisi arbitrairement. Le critère naturellement s'applique avec les mêmes propriétés au cas général de sélection parmi toutes les règles d'association.

Le tableau 1 liste les 12 premières règles rangées par valeurs décroissantes du critère VT100. Considérons par exemple la première règle :

> Le conséquent C est l'item « Less US$ 50 » avec le support n[C] = 11 221
> L'antécédent A est « Marital status=never married » avec le support n[A] = 4 918
> 4 693 individus satisfont à cette règle (soit un support de 32%)

On considère l'hypothèse d'indépendance entre l'antécédent et le conséquent exprimée par la loi hypergéométrique de paramètres n=14 734, n[A] = 4 918 et n[C] = 11 221. Sous cette hypothèse d'indépendance, on trouvera que la probabilité d'un événement au moins aussi extrême que *« observer 4 693 individus ou davantage satisfaisant à la règle »* est égale à la probabilité de se trouver au-delà de 38,9 écarts types pour une loi normale centrée réduite. C'est la valeur du critère *Valeur-test VT* lue dans la colonne 10 du tableau.

Le critère VT100 est consigné dans la colonne 8 du tableau. Il indique que, si on ramenait la taille de l'échantillon observée n = 14 743 à la taille *arbitraire* n = 100, la probabilité d'un événement au moins aussi extrême, transposé dans ce contexte, serait égale à la probabilité d'être au delà de 3,2 écarts types de la loi normale (soit une *p-value* inférieure à 0,0007 calculée sur un échantillon de taille 100). Ce serait donc bien un événement *exceptionnel* qui nous ferait douter de l'indépendance entre l'antécédent et le conséquent, marquant par-là

l'intérêt que présente cette règle. Dans ce tableau 1, les 7 premières règles correspondent à des événements dont la probabilité calculée sur un échantillon de taille 100 serait inférieure à 0,001. La dernière des 12 règles du tableau aurait elle-même une probabilité légèrement inférieure à 0,05.

N°	Antécédent	n[A]	n[A^C]	Support	Confiance	Lift	VT 100	VT100 simul	VT
1	marital_status=**Never-married**	4 918	4 693	0,32	0,95	1,25	3,20	3,01	38,9
2	marital_status=**Never-married** - "native_country=**United-States**"	4 524	4 318	0,29	0,95	1,25	3,00	2,83	36,6
3	workclass=**Private** - "marital_status=**Never-married**"	4 104	3 947	0,27	0,96	1,26	2,91	2,61	35,5
4	workclass=**Private** - "marital_status=**Never-married**" - "native_country=**United-States**"	3 748	3 608	0,24	0,96	1,26	2,74	2,42	33,5
5	marital_status=**Never-married** - "race=**White**"	4 045	3 839	0,26	0,95	1,25	2,65	2,46	32,9
6	marital_status=**Never-married** - "race=**White**" - "native_country=**United-States**"	3 798	3 605	0,24	0,95	1,25	2,53	2,28	31,5
7	workclass=**Private** - "marital_status=**Never-married**" - "race=**White**"	3 379	3 237	0,22	0,96	1,26	2,43	2,13	30,5
8	sex=**Female**	4 972	4 423	0,30	0,89	1,17	1,96	1,92	26,1
9	workclass=**Private** - "sex=**Female**"	3 949	3 582	0,24	0,91	1,19	1,89	1,79	25,1
10	sex=**Female** - "native_country=**United-States**"	4 582	4 068	0,28	0,89	1,17	1,80	1,76	24,2
11	workclass=**Private** - "sex=**Female**" - "native_country=**United-States**"	3 611	3 267	0,22	0,90	1,19	1,72	1,67	23,3
12	relationship=**Not-in-family**	3 833	3 431	0,23	0,90	1,18	1,65	1,52	22,6

TAB. 1 -- *Les 12 premières règles calculées sur les 14 743 individus Le support du conséquent est n[C] = 11 221*

Quand il faut limiter la liste des règles fournies par les algorithmes de recherche de règles, on peut choisir de les ranger en fonction du critère VT100 et fixer un seuil d'arrêt. Par analogie avec les seuils conventionnels adoptés par les statisticiens, on pourra s'arrêter à la valeur 1,645 du critère VT100 (correspondant au seuil 0,05 pour une p-value) ou à la valeur 2,326 (seuil 0,01 pour une p-value) ou encore 3,09 (seuil 0,001) dans les cas où il y aurait profusion de règles intéressantes.

Dans la colonne 9 du tableau 1, on fait figurer la valeur du critère VT100 estimée par simulation sur des échantillons réels de taille 100 extraits du tableau des observations. La procédure de simulation adoptée ici est la suivante : nous réalisons un tirage aléatoire avec remise parmi les individus couverts par la règle puis nous calculons la valeur test VT correspondante ; cette procédure est répétée N fois (N = 50) et la valeur test calculée par simulation est la moyenne arithmétique des valeurs test individuelles.

On remarque que la VT100 calculée par interpolation à partir de la loi hypergéométrique a tendance à surestimer la valeur obtenue par simulation sur des échantillons de taille 100.

5.2 Validation avec un échantillon test

Pour compléter cet exemple, on a procédé à une approche de validation sur échantillon test. Les 14 743 individus ont été répartis au hasard par moitié dans un échantillon d'apprentissage (7 371 cas) et un échantillon test (7 372 cas). Les règles ont été recalculées sur l'échantillon d'apprentissage. Dans le tableau 2-A, on liste les règles obtenues en les numérotant en colonne 1 avec le numéro qu'on leur a attribué dans le tableau 1. On constate que 9 des 12 règles se retrouvent avec l'échantillon d'apprentissage. L'élimination au hasard de la moitié des cas a entraîné la disparition des règles n°4 et n°7 qui étaient présentes dans le tableau 1 et n'a pas fait apparaître des règles nouvelles en haut du classement par les VT100. On constate dans le bas du tableau deux inversions de classement par les VT100 dont les valeurs numériques restent cependant proches.

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
N°	Antécédent	n[A]	n[A^C]	Support	Con-fiance	Lift	VT 100	VT
1	marital_status=**Never-married**	2 502	2 398	0,33	0,96	1,25	3,26	27,9
2	marital_status=**Never-married** - "native_country=**United-States**"	2 314	2 222	0,30	0,96	1,25	3,11	26,6
3	workclass=**Private** - "marital_status=**Never-married**"	2 083	2 009	0,27	0,96	1,26	2,95	25,2
5	marital_status=**Never-married** - "race=**White**"	2 031	1 939	0,26	0,95	1,25	2,71	23,6
6	marital_status=**Never-married** - "race=**White**" - "native_country=**United-States**"	1 930	1 844	0,25	0,96	1,25	2,60	22,8
8	sex=**Female**	2 502	2 235	0,30	0,89	1,17	1,97	18,5
9	workclass=**Private** - "sex=**Female**"	1 990	1 805	0,24	0,91	1,18	1,83	17,4
10	sex=**Female** - "native_country=**United-States**"	2 315	2 068	0,28	0,89	1,17	1,84	17,4
11	workclass=**Private** - "sex=**Female**" - "native_country=**United-States**"	1 829	1 657	0,22	0,91	1,18	1,69	16,3
12	relationship=**Not-in-family**	1 898	1 715	0,23	0,90	1,18	1,71	16,4

TAB. 2-A **--** *Echantillon d'apprentissage (7 341 individus), le support du conséquent est n[C] = 5 648*

L'application au fichier test des 10 règles trouvées sur le fichier d'apprentissage conduit aux résultats du tableau 2-B. Les règles sélectionnées avec le critère VT100 appliqué à l'échantillon test, avec le même seuil 1,645 (seuil 0,05 d'une p-value), constituent ici la même liste de 10 règles, rangées dans le même ordre que dans le tableau 1.

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
N°	Antécédent	n[A]	n[A^C]	Support	Con-fiance	Lift	VT 100	VT
1	marital_status=**Never-married**	2 416	2 295	0,31	0,95	1,26	3,13	27,0
2	marital_status=**Never-married** - "native_country=**United-States**"	2 210	2 096	0,28	0,95	1,25	2,89	25,1
3	workclass=**Private** - "marital_status=**Never-married**"	2 021	1 938	0,26	0,96	1,27	2,89	24,9
5	marital_status=**Never-married** - "race=**White**"	2 014	1 900	0,26	0,94	1,25	2,60	22,9
6	marital_status=**Never-married** - "race=**White**" - "native_country=**United-States**"	1 868	1 761	0,24	0,94	1,25	2,43	21,7
8	sex=**Female**	2 470	2 188	0,30	0,89	1,17	1,96	18,4
9	workclass=**Private** - "sex=**Female**"	1 959	1 777	0,24	0,91	1,20	1,94	18,1
10	sex=**Female** - "native_country=**United-States**"	2 267	2 000	0,27	0,88	1,17	1,75	16,8
11	workclass=**Private** - "sex=**Female**" - "native_country=**United-States**"	1 782	1 610	0,22	0,90	1,19	1,73	16,6
12	relationship=**Not-in-family**	1 935	1 716	0,23	0,89	1,17	1,60	15,5

TAB. 2-B **--** *Echantillon test (7 342 individus), le support du conséquent est n[C] = 5 575*

Il est intéressant finalement de comparer, pour ces 10 règles, le critère VT100 évalué sur l'échantillon total, sur l'échantillon d'apprentissage et sur l'échantillon test. Les résultats qui apparaissent sur le tableau 3 montrent bien que le critère est surévalué sur l'échantillon d'apprentissage (phénomène classique) et qu'il établit un compromis (sorte de moyenne) avec la valeur calculée sur la totalité des cas disponibles.

[1]	[2]	[3]	[4]	[5]
N°	Antécédent	VT 100 total (100%)	VT 100 (50% appr)	VT 100 (50% test)
1	marital_status=**Never-married**	3,20	3,26	3,13
2	marital_status=**Never-married** - "native_country=**United-States**"	3,00	3,11	2,89
3	workclass=**Private** - "marital_status=**Never-married**"	2,91	2,95	2,89
5	marital_status=**Never-married** - "race=**White**"	2,65	2,71	2,60
6	marital_status=**Never-married** - "race=**White**" - "native_country=**United-States**"	2,53	2,60	2,43
8	sex=**Female**	1,96	1,97	1,96
9	workclass=**Private** - "sex=**Female**"	1,89	1,83	1,75
10	sex=**Female** - "native_country=**United-States**"	1,80	1,84	1,94
11	workclass=**Private** - "sex=**Female**" - "native_country=**United-States**"	1,72	1,69	1,73
12	relationship=**Not-in-family**	1,65	1,71	1,60

TAB. 3 -- Comparaison du critère VT100 appliqué à l'ensemble des observations, à l'échantillon d'apprentissage et à l'échantillon test

6 Conclusion

Le critère VT100 est proposé pour ranger et sélectionner un nombre raisonnable de règles d'association dans les cas d'applications réelles où les données à analyser sont volumineuses.

Ce critère présente des propriétés intéressantes :

- Facile à comprendre puisqu'on s'appuie sur le mécanisme usuel des tests en raisonnant sur la population des échantillons de taille 100 extraits des observations.
- Facile à calculer quelle que soit la taille des données (car on effectue quelques interpolations dans la table hypergéométrique limitée à n=100).
- Souple puisque ne dépendant que d'un seuil qui peut s'exprimer en terme de p-value (α=0,05 ou α=0,01 etc.) ou en terme de nombre d'écarts types d'une loi normale (valeur-test 1,645 ou 2,326 etc.).
- Indépendant de la taille des données tout en faisant sur les informations (support, confiance, lift, etc.) les compromis que savent faire les critères statistiques classiques (qui, appliqués directement, sont très sensibles à la taille de l'échantillon).

Si l'on veut rendre plus robuste la sélection des règles (quand le nombre de cas disponibles est élevé) on propose la stratégie suivante : on divisera par moitié les données en apprentissage et test. On calculera les règles sur l'échantillon d'apprentissage et on évaluera le critère VT100 sur l'échantillon test. On rangera les règles et on les sélectionnera par seuil en fonction du critère VT100 évalué sur l'échantillon test.

Enfin, notons que le critère VT100 peut être mis en application dans de nombreux autres problèmes de sélection *d'items caractéristiques* rencontrés dans le Data mining : par exemple le critère de choix des variables de coupure dans un arbre de segmentation ou le critère d'élagage associé ; le classement des items caractérisant une classe dans une typologie, ou caractérisant un facteur dans une analyse factorielle, etc.

Références

Agrawal, R., R. Srikant (1994). Fast Algorithms for Mining Association Rules, in *Proc. of 20th International Conf. On Very Large Databases (VLDB)*, 478-499.

Gras, R., R. Couturier, M. Bernadet, J. Blanchard, H. Briand, F. Guillet, P. Kuntz, R. Lehn, P. Peter (2002), Quelques critères pour une mesure de qualités de règles d'association – Un exemple : l'intensité d'implication, *Rapport de recherche pour le groupe de travail GAFOQUALITE de l'action spécifique STIC fouille de bases de données*, Ecole Polytechnique de Nantes.

Lallich, S., O. Teytaud (2004), Evaluation et validation de l'intérêt des règles d'association, *Revue des Nouvelles Technologies de l'Information*, RNTI-E-1:193-218.

Lebart, L., A. Morineau, M. Piron (1995). *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, S. Sallich (2003). Critères d'évaluation des mesures de qualité en ECD, *Revue des Nouvelles Technologies de l'Information*, RNTI-1:123-134.

Lerman, I.C., J. Azé (2003), Une mesure probabiliste contextuelle discriminante de qualité des règles d'association, *EGC'2003*, 1(17):247-262.

Morineau, A. (1984). Note sur la caractérisation statistique d'une classe et les valeurs-tests, *Bull. Techn. du Centre de Statist. et d'Infor. Appl.,* 2:20-27.

Newman, D.J., S. Hettich, C.L. Blake, C.J. Merz (1998). UCI Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Rakotomalala, R. (2005). TANAGRA : Un logiciel gratuit pour l'enseignement et la recherche, *EGC'2005*, 697-702, http://chirouble.univ-lyon2.fr/~ricco/tanagra.

Tan, P.N., V. Kumar, J. Srivastava (2002). Selecting the right interestingness measure for association patterns, in *Proceedings of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32-41.

Summary

Extraction of association rules from databases often generates a large number of rules. To rank and validate the rules, many statistical measurements have been proposed. They allow to highlight such or such characteristics of the extracted rules. They share the property to be increasing with database size and very often lead to the acceptance of nearly all the rules when applied to huge data sets. We propose a new criteria derived from the concept of test-value. It presents as principal characteristic to be insensitive to the database size, thus avoiding the pitfall of fallaciously relevant rules. It also makes possible to compare rules extracted from different databases. It also makes possible to manage various thresholds of significance for the rules. The behavior of this criteria is detailed on an example.