Contrôle du risque multiple pour la sélection de règles d'association significatives

Stéphane Lallich*, Elie Prudhomme*, Olivier Teytaud**

*Laboratoire E.R.I.C, Université Lumière Lyon 2
5, avenue Pierre Mendès-France, 69676 BRON Cedex – France
stephane.lallich@univ-lyon2.fr, Elie.Prudhomme@etu.univ-lyon2.fr

**Artelys
215 avenue Jean-Jacques Rousseau, 92136 Issy-les-Moulineaux
olivier.teytaud@artelys.com

Résumé. Les algorithmes d'extraction de règles d'association parcourent efficacement le treillis des itemsets pour constituer une base de règles admissibles à des seuils de support et de confiance, mais donnent une multitude de règles peu exploitables. Nous suggérons d'épurer de telles bases en éliminant les règles non statistiquement significatives. La multitude de tests pratiqués conduit mécaniquement à multiplier les règles sélectionnées à tort. Après avoir présenté des procédures issues de la biostatistique qui contrôlent non pas le risque, mais le nombre de fausses découvertes, nous proposons BS_FD, un algorithme original fondé sur le bootstrap qui sélectionne les règles significatives en contrôlant le nombre de fausses découvertes. Des expérimentations montrent l'efficacité de ces procédures.
Mots-clefs : Règle d'association, qualité, contrôle du risque multiple.

1 Admissibilité, intérêt et signification statistique

La recherche des règles d'association intéressantes est un problème classique de l'Extraction des Connaissances à partir des Données à la suite des travaux de [Agrawal et al., 1993] dans le cadre des bases de données transactionnelles. Dans une telle base, un enregistrement est une transaction et les champs correspondent aux articles disponibles. On note n le nombre de transactions et p le nombre d'articles. L'acte d'achat (item) associé à chaque article est une variable booléenne. Sur l'ensemble des transactions, on a une matrice booléenne X, de dimensions n et p. La conjonction des actes d'achat (*itemset*) associés à un ensemble d'articles est vue comme une variable booléenne.

A partir de la matrice booléenne X, on veut extraire des règles du type "si un client achète du pain et du fromage, alors probablement il achète aussi du vin". Une règle d'association est une expression r du type $A \rightarrow B$, où l'antécédent A et le conséquent B sont des itemsets qui n'ont pas d'items communs. On note n_a et n_b les nombres de transactions qui réalisent respectivement les items de A et de B, n_{ab} le nombre de celles qui réalisent à la fois A et B. Les proportions correspondantes sont désignées par p_a, p_b et p_{ab}. Ce formalisme se généralise à toute base de données dont on a extrait une table booléenne cas-attributs.

Les algorithmes d'extraction usuels reposent sur le support et la confiance, en particulier *Apriori* l'algorithme fondateur [Agrawal et Srikant, 1994] et les améliorations qui en ont été proposées. Le support d'une règle est la proportion de transactions qui réalisent à la fois A et B, $Supp(A \rightarrow B) = p_{ab} = \frac{n_{ab}}{n}$, alors que sa confiance est la proportion de transactions qui réalisent B, parmi celles qui réalisent A, $Conf(A \rightarrow B) = \frac{p_{ab}}{p_a} = \frac{n_{ab}}{n_a} = 1 - \frac{n_{a\bar{b}}}{n_a}$.

Les algorithmes d'extraction "support-confiance" parcourent le treillis des *itemsets* pour rechercher les *itemsets* fréquents, dont le support dépasse un seuil min_{supp}, avec une efficacité liée à l'antimonotonie du treillis. On en déduit les règles dont la confiance dépasse le seuil min_{conf}, obtenant la base de règles admissibles aux seuils choisis. De telles bases comportent un grand nombre de règles, pas toujours intéressantes.

La sélection des règles intéressantes à partir d'une base de règles admissibles nécessite d'évaluer celles-ci à l'aide de mesures ayant les qualités requises compte tenu de la nature des règles d'association et des attentes de l'utilisateur. Nous avons recensé de telles mesures et proposé des critères pour les évaluer [Lallich et Teytaud 2003], ainsi qu'une procédure d'aide à la décision pour les choisir [Lenca et al., 2003]. Pour chaque mesure choisie, on fixe le seuil minimal à partir duquel une règle est sélectionnée ou l'on retient un nombre fixé des meilleures règles.

Le critère "prise en compte du nombre d'observations" oppose les mesures statistiques et les mesures descriptives. Il est logique *a priori* de souhaiter qu'une mesure soit statistique, les résultats observés étant d'autant plus fiables que n est grand. Cependant, compte tenu de la taille des bases sur lesquelles on recherche des règles d'association, de telles mesures perdent leur pouvoir discriminant, ainsi l'indice d'implication [Lerman et al., 1981] et l'intensité d'implication [Gras 1979]. Des solutions très intéressantes ont été proposées : l'indice probabiliste discriminant [Lerman et Azé, 2003] qui centre et réduit les valeurs de l'indice d'implication sur une base de règles et l'intensité d'implication entropique [Gras et al., 2001] qui affecte l'intensité d'implication d'un facteur correctif tenant compte de l'entropie des expériences B/A et $\overline{A}/\overline{B}$. Mais il s'ensuit un mélange des notions de signification et d'intérêt, une perte d'intelligibilité de la mesure et sa loi est plus difficile à étudier.

Paradoxalement, le nombre de cas n est le même pour toutes les règles de la base, ce qui milite pour d'abord tester la signification statistique des règles, au sens de savoir si elles renforcent réellement la probabilité du conséquent. On testera l'hypothèse d'indépendance de A et B, notée H_0, en direction d'une dépendance positive (hypothèse alternative unilatérale notée H_1), pour ensuite utiliser des mesures descriptives intelligibles et discriminantes de l'intérêt des règles sur la base filtrée. Nous proposons ainsi une nouvelle démarche qui dissocie la signification statistique de l'évaluation de l'intérêt. Elle comporte trois étapes :

- étape 1 : application d'un algorithme "support-confiance" à la base de cas, pour constituer une base de règles admissibles aux seuils choisis.
- étape 2 : filtrage de la base de règles par le test d'indépendance de A et B pour chaque règle, soit une multitude de tests ; on pourrait tester tout à la fois l'indépendance et le dépassement significatif des seuils de support et de confiance.
- étape 3 : analyse des règles figurant dans la base filtrée par des mesures descriptives vérifiant les critères retenus par l'utilisateur [Lenca et al., 2003].

Dans cet article, nous approfondissons l'étape 2 qui pose le problème de contrôler la multiplicité de tests, pour éviter l'inflation de "faux positifs". En effet, si chaque test est pratiqué au risque de 1^e espèce α, on engendre mécaniquement des "faux positifs", ici des règles sélectionnées alors qu'elles ne renforcent pas réellement la probabilité du conséquent. Nous avons proposé des méthodes de contrôle du risque utilisant la théorie de l'apprentissage statistique et la *VC-dimension* [Teytaud et Lallich, 2001], ou le *bootstrap* [Lallich et Teytaud, 2003]. Dans la pratique, ces méthodes sont peu puissantes, ignorant des règles significatives.

La section 2 détaille le test de signification appliqué à chaque règle. Le contrôle des faux positifs et l'idée de contrôler le nombre de fausses découvertes plutôt que le risque sont exposés en section 3. Les procédures de contrôles récemment développées en biostatistique sont présentées en section 4 et une méthode de contrôle originale fondée sur le bootstrap est proposée en section 5. Enfin, nous appliquons ces méthodes pour filtrer quelques bases de règles (section 6) et nous concluons (section 7).

2 Test de signification d'une règle

Considérons une règle $A \rightarrow B$ et une mesure de qualité μ, croissante avec n_{ab} à marges fixées. On dira que la règle $A \rightarrow B$ est significative au sens de la mesure μ si la valeur $\mu_{obs} = \mu(A \rightarrow B)$ remet en cause H_0 dans le sens de H_1, au risque de 1^e espèce α. Pour décider de la signification, on calcule la *p-value* de μ_{obs} qui est la probabilité d'obtenir une valeur aussi grande que μ_{obs} sous H_0 et on sélectionne la règle si l'on a *p-value* $< \alpha$. Cette démarche impose de connaître la loi de $\mu(A \rightarrow B)$ sous H_0.

Dans le cadre d'une modélisation à marges fixées, on suppose que sous H_0 les "1" de A et les "1" de B sont répartis au hasard, indépendamment, en respectant les marges. On démontre que le nombre d'exemples suit une loi hypergéométrique, $N_{ab} \equiv H(n,np_a,p_b) \equiv H(n,np_b,p_a)$ (par convention, les variables aléatoires sont en majuscules). Pour une valeur observée n_{ab}, à marges fixées :

$p - value = \Pr(N_{ab} \geq n_{ab}) = \Pr(H(n,np_a,p_b) \geq n_{ab})$

On peut opérer l'approximation normale de cette loi hypergéométrique sous des conditions peu contraignantes, à savoir $n_a n_b \geq 5n$ et $n_a n_{\overline{b}} \geq 5n$, où Φ est la fonction de répartition de la loi normale centrée réduite $N(0,1)$. En notant $t_{ab} = \frac{n_{ab}}{n_a n_b}$, la valeur attendue de N_{ab} sous H_0, r le coefficient de corrélation entre A et B, il vient :

$$p - value = 1 - \Phi\left(\frac{n_{ab} - t_{ab}}{\sqrt{\frac{n_{\overline{a}}}{n-1} t_{ab} p_{\overline{b}}}}\right) \approxeq 1 - \Phi\left(\frac{n_{ab} - t_{ab}}{\sqrt{n p_a p_b p_{\overline{a}} p_{\overline{b}}}}\right) = 1 - \Phi\left(r\sqrt{n}\right)$$

Le coefficient de corrélation r est donc une mesure privilégiée pour tester l'indépendance (H_0) face à une dépendance positive (H_1).

3 Risque et erreurs de 1^e espèce

Pour rechercher les règles $A \rightarrow B$ statistiquement significatives parmi les m règles de la base de règles, on répète m fois le test d'indépendance entre A et B face à une dépendance positive. On rencontre ainsi un problème classique en fouille des données,

Réalité \ Décision	Acceptation	Rejet	Total
H_0 est vraie	U	V	m_0
H_1 est vraie	T	S	m_1
Total	W	R	m

TAB. 1 – *Synthèse des résultats de m tests*

le contrôle des erreurs de 1^e espèce (ou faux positifs). Si l'on teste m règles non significatives au risque α, mécaniquement on sélectionne à tort $m\alpha$ règles (soit 500 règles, si $m = 10000$ et $\alpha = 0.05$). La correction de Bonferroni, qui consiste à pratiquer chaque test au risque $\frac{\alpha}{m}$, pour que le risque de rejeter au moins une fois à tort H_0, noté par la suite $FWER$, soit égal à α, n'est pas une bonne solution pour deux raisons :

- le $FWER$ est en fait non contrôlé, compris entre $\frac{\alpha}{m}$et α, ne valant α que si toutes les règles sont indépendantes.
- le $FWER$ est très conservateur en faveur de H_0, ce qui augmente considérablement le risque de 2^e espèce ou risque de ne pas sélectionner une règle pertinente.

Pour régler ce problème, il faut évaluer les erreurs de 1^e espèce par une quantité moins sévère que le $FWER$ et contrôler celle-ci, notamment lorsque les tests ne sont pas indépendants. Les règles ne sont pas indépendantes en raison des items qu'elles partagent et des dépendances entre items. Différentes solutions ont été développées, le plus récemment en sélection de gènes s'exprimant différemment suivant l'étiquette de la biopsie. On trouvera une remarquable synthèse de ces travaux dans [Ge et al., 2003].

L'idée fondamentale [Benjamini et Hochberg, 1995] est de considérer non pas le risque d'erreur de la procédure de test lorsqu'elle est pratiquée une fois, mais le nombre d'erreurs commises lorsque l'on réitère m fois la procédure. A partir du tableau 1 (où les quantités en majuscules sont des variables aléatoires observables, celles en minuscules étant fixes, mais inconnues en ce qui concerne m_0 et m_1), on peut définir différents indicateurs des erreurs commises. Nous présentons ici les deux plus connus, le $FWER$ (*Family wise error rate* ou taux d'erreur en famille complète) et le FDR (*False discovery rate* ou taux de fausses découvertes).

$FWER$ est la probabilité de rejeter au moins une fois à tort H_0, $FWER = \Pr(V > 0)$. Il a l'inconvénient d'être bien trop sévère pour une multiplicité de tests, mais il nous a suggéré une variante flexible originale, où l'on s'autorise V_0 fausses découvertes. Nous l'appelons *User Adjusted Family Wise Error Rate*, $UAFWER = \Pr(V > V_0)$, pour le contrôle duquel nous proposons un algorithme fondé sur le bootstrap (section 5).

Pour remédier aux inconvénients du $FWER$, diverses quantités reposant sur l'espérance de V, le nombre de fausses découvertes, éventuellement normalisée, ont été proposées. La plus connue est le FDR [Benjamini et Hochberg 1995], la proportion attendue de règles sélectionnées à tort parmi les règles sélectionnées. Lorsque $R = 0$, on pose $\frac{V}{R} = 0$, soit $FDR = E(Q)$, où Q vaut $\frac{V}{R}$ si $R > 0$, 0 sinon. Il s'ensuit :

$FDR = E(\frac{V}{R} \mid R > 0)P(R > 0)$

[Storey, 2001] a proposé le *pFDR,* une variante du *FDR,* adaptée à l'estimation du taux d'erreur sachant que l'hypothèse nulle a été refusée au moins une fois :

$pFDR = E(\frac{V}{R} \mid R > 0)$

Ces quantités ont l'intérêt d'être moins sévères sur le résultat des m tests, au prix de l'acceptation de quelques sélections à tort dont on contrôle la proportion, ce qui augmente pour chaque test la probabilité pour qu'une règle pertinente soit sélectionnée (puissance). On a $FDR \leq FWER$ et $FDR \leq pFDR$, d'où $FDR \leq pFDR \leq FWER$ pour m grand, car $Pr(R > 0)$ tend vers 1 quand m croît. Une fois choisie une définition du risque multiple de 1^e espèce, se pose le problème de son contrôle. Nous allons examiner successivement le cas du *FWER* et celui du *FDR*.

4 Procédures de contrôle

4.1 Contrôle du *FWER*

Correction de Bonferroni La correction de Bonferroni consiste à calculer des *p-values* ajustées afin de prendre en compte la multiplicité des tests. Etant données la statistique de test T_r relative à la règle r, $r = 1,2,...m$, et la *p-value* correspondante p_r, on définit la *p-value* ajustée, notée $\widetilde{p}_r$, par $\widetilde{p}_r = \min\{mp_r,1\}$. On sélectionne toutes les règles ayant une *p-value* ajustée inférieure au risque α_0. On montre :

$FWER = 1 - \Pr(\bigcap_{r=1}^{m} \left(P_r > \frac{\alpha_0}{m}\right) \mid H_0)$

Sous condition d'indépendance des règles, il vient $FWER = 1 - (1 - \frac{\alpha_0}{m})^m \approx \alpha_0$. A défaut d'indépendance, on a : $\frac{\alpha_0}{m} \leq FWER \leq \alpha_0$.

Procédure Step-down de Holm Les procédures pas à pas examinent les *p-values* par ordre croissant et font évoluer le seuil tout au long de la procédure. [Holm, 1979] considère qu'une variable sélectionnée correspond à une situation où H_0 est fausse, ce qui amène à ne prendre en compte pour le nouveau seuil que les variables restant à examiner. Les *p-values* étant rangées dans l'ordre croissant, où $p_{(k)}$ désigne la k^e *p-value*, on refuse H_0 tant que $p_{(k)} < \frac{\alpha_0}{m-k+1}$. On accepte H_0 pour toutes les *p-values* qui suivent la première acceptation. Cette procédure, facile à mettre en oeuvre, donne de bons résultats lorsque le nombre de tests est faible, la correction du seuil ayant alors de l'importance. Elle reste mal adaptée à un grand nombre de tests.

Procédure minP de Westfall et Young [Westfall et Young, 1993] ont proposé minP une procédure d'ajustement des *p-values* qui contrôle le FWER mais oblige à calculer $\widetilde{p}_r = \Pr(\min_{k=1,2,...,m} P_k \leq p_r \mid H_0)$. Ce calcul doit être opéré par randomisation des étiquettes de cas, si les variables (ici des règles) ne sont pas indépendantes. Bien adapté à la recherche des gènes s'exprimant différemment suivant l'étiquette de la biopsie, ce procédé ne convient pas à la recherche des règles d'association.

4.2 Contrôle du *FDR*

Procédure Benjamini, Liu On doit à [Benjamini et Liu, 1999], une méthode séquentielle pour contrôler le *FDR* en cas d'indépendance. Les *p-values* sont prises dans l'ordre croissant et l'on rejette l'hypothèse nulle tant que la *p-value* examinée $p_{(i)}$ est inférieure à $\frac{i\alpha_0}{m}$. Cette procédure assure un *FDR* égal à $\frac{m_0}{m}\alpha_0$ en cas d'indépendance. Elle est compatible avec des données positivement dépendantes.

Procédure SAM de Storey Cette procédure [Storey, 2001] repose sur l'estimation de m_0 puis celle de $E(V)$ et enfin celle du *FDR*. Nécessitant une randomisation des étiquettes, elle n'est pas adaptée au cas des règles.

pFDR Pour estimer $pFDR = E(\frac{V}{R} \mid R > 0)$, qui est la proportion de fausses détections, on utilise l'approximation [Storey, 2001] :

$\hat{pFDR}(\delta) = \frac{\hat{\pi_0}.m.\delta}{\#\{p_i \leq \delta, i=1,\ldots,m\}}$

- m est le nombre de variables à tester, ici le nombre de règles ;

- δ définit la zone de rejet ; les hypothèses correspondant aux p-value inférieures ou égales à δ sont rejetées ;

- p_i est la i^{eme} plus grande *p-value* ;

- $\pi_0 = \frac{m_0}{m}$ est la proportion d'hypothèses nulles ; ici, π_0 est estimé par $\hat{f}(1)$, où $\hat{f}$ est une *cubic spline* avec 3 degrés de liberté de $\hat{\pi_0}(\lambda)$ sur λ. On a $\hat{\pi_0}(\lambda) = \frac{\#\{p_i \geq \lambda, i=1,\ldots,m\}}{m(1-\lambda)}$ et λ désigne la zone d'acceptation dont les valeurs sont comprises entre 0 et 0.95.

Le *pFDR* est donc défini par rapport à une zone de rejet qu'il faut choisir par avance. Une fois le *pFDR* global calculé, les variables sont contrôlées par une procédure *step-down* grâce aux *q-values* définies pour chaque *p-value* par $\hat{q}(p_m) = \hat{\pi_0}.p_m$ et :

$$\hat{q}(p_i) = \min\left(\frac{\hat{\pi_0}.m.p_i}{i}, \hat{q}(p_{i+1})\right); i = m-1, \ldots, 1$$

La *q-value* est au *pFDR* ce que la *p-value* est à l'erreur de 1^e espèce ou ce que la *p-value* ajustée est au *FWER*. Pour toute règle dont la *p-value* a une *q-value* inférieure au *pFDR,* on rejette H_0 et l'on sélectionne la règle.

5 Contrôle du *UAFWER* par l'algorithme BS_FD

5.1 Notations

- on note $\mathcal{T}$ l'ensemble des transactions, $n = Card(\mathcal{T})$, p le nombre d'items ;
- $\mathcal{R}$ une base de règles d'association valides au sens de critères prédéfinis, par exemple le support et la confiance, $m = Card(\mathcal{R})$, $\mathcal{R}^*$ un sous-ensemble de $\mathcal{R}$ rassemblant les règles valides significatives au sens du critère c ;
- $c(r)$ désigne l'évaluation de la règle r selon le critère c, $c'(r)$ l'évaluation empirique de la règle r selon le critère c sur l'ensemble T ;
- V : nombre de faux positifs, δ : risque de la procédure de contrôle des faux positifs, avec V_0 nombre de faux positifs que l'on ne souhaite pas dépasser au risque δ.

5.2 But

On veut sélectionner parmi les règles r de la base $\mathcal{R}$ celles qui sont statistiquement significatives pour le critère c, au sens où leur évaluation $c(r)$ est significativement plus élevée que $c_0(r)$, valeur attendue sous H_0, l'hypothèse d'indépendance de A et B.

Nous avons suggéré [Lallich et Teytaud, 2003] différents algorithmes qui utilisent les outils de l'apprentissage statistique pour garantir que 100% des règles trouvées sont significatives au risque α donné, ainsi l'algorithme *BS* fondé sur le bootstrap.

Les expérimentations ont confirmé que cette approche était trop prudente et par là même peu puissante. Prenant en compte l'idée d'accepter de façon contrôlée un certain nombre de fausses découvertes, à l'instar des travaux de Benjamini (section 3), nous proposons *BS_FD* qui adapte l'algorithme *BS* au contrôle du nombre de faux positifs.

L'algorithme *BS_FD* sélectionne les règles candidates de telle sorte que l'on contrôle $UAFWER = P(V > V_0)$, assurant que le nombre de règles sélectionnées à tort (faux positif) ne dépasse pas V_0 au risque δ. Plus précisément, on garantit que $P(V > V_0)$ converge vers δ, à la limite d'un grand échantillon de transactions.

5.3 Algorithme *BS_FD*

Pour garantir $c(r) > c_0(r)$, on peut se limiter sans perte de généralité à garantir $c(r) > 0$, en remplaçant $c(r)$ par le critère translaté $c(r) - c_0(r)$. On note par la suite # l'opérateur "cardinal", associant à un ensemble son cardinal (i.e. $\#E$ est le cardinal de l'ensemble E.

1. Définir $c'(r)$, l'évaluation empirique de $c(r)$ sur l'ensemble $\mathcal{T}$ de transactions.
2. Un grand nombre de fois (pour i = 1,2,...,l) - Tirer au sort, avec remise, une liste $\mathcal{T}'$ d'éléments de $\mathcal{T}$, de même cardinal que $\mathcal{T}$.
   - Définir $c"(r)$, l'évaluation de $c(r)$ sur la liste T' de transactions.
   - Calculer $\varepsilon(V_0,i)$ minimal, tel que $\#\{c"(r) > c'(r) + \varepsilon(V_0,i)\} \leq V_0$
3. On obtient l valeurs $\varepsilon(V_0,i)$. Calculer $\varepsilon(\delta)$, le quantile $(1-\delta)$ des $\varepsilon(V_0,i)$.
4. Garder dans $\mathcal{R}^*$ toutes les règles r de $\mathcal{R}$ telles que $c\prime(r) > \varepsilon(\delta)$.

Pour réaliser la dernière partie de l'étape 2. de l'algorithme, on applique la procédure "calculerEpsilon(i,δ,V_0)" définie ci-dessous.

Procédure "calculerEpsilon(i,δ,V_0)**"**
- *tableauEpsilon* = (0,0,...,0) tableau de taille m
- pour r variant de 0 à $m-1$: *tableauEpsilon*$(r) = c"(r) - c'(r)$
- Calculer le $(V_0+1)^e$ plus grand élément de *tableauEpsilon*

5.4 Justification de la méthode

Les méthodes de bootstrap [Efron, 1979] ont d'abord un aspect très intuitif de par leur idée d'approcher l'écart entre la loi empirique et la loi réelle par l'écart entre la loi bootstrappée et la loi empirique. En outre, elles ont de profondes justifications mathématiques [Giné et Zinn, 1984] qui nécessitent une formalisation précise de la question posée.

Formellement, l'objectif est que la fonction de répartition du nombre de règles telles que $c(r) < 0$ malgré $c'(r) > \epsilon$ ait pour valeur au moins $1-\delta$ en V_0. On a $\#\{c(r) \leq 0$ et $c'(r) > \epsilon\}$ majoré par $\#\{c'(r) \geq c(r) + \epsilon\}$.

Les théorèmes sur le bootstrap appliqué à une famille de fonctions vérifiant des hypothèses minimales [Van der Waart et Wellner, 1996], nous permettent d'approcher cette quantité par $\#\{c''(r) \geq c'(r) + \epsilon\}$.

5.5 Cas de plusieurs critères

En pratique, on s'intéresse souvent à plusieurs critères, dans notre cas à la confiance, au support, plus un critère de non-indépendance. L'extension de l'algorithme *BS_FD,* notée *BS_FD_mc* se fait simplement en utilisant comme critère unique le min des différents critères. Ainsi pour un travail sur 3 critères c_1, c_2 et c_3, considère-t-on $c(r) = min\{c_1(r),c_2(r),c_3(r)\}$. Utiliser *BS_FD_mc* sur c au risque δ fournit bien des règles garanties à la fois pour les critères c_1, c_2 et c_3 au risque δ.

Afin d'optimiser le risque de seconde espèce, on gagnera à travailler sur des transformations (différentiables au sens de Hadamard) des c_i qui rendent ces critères homogènes, par exemple des *p-values* ou des réductions, en divisant l'écart empirique de chaque critère à sa valeur de référence par l'estimation de l'écart-type issue des $c\prime(r)$.

5.6 Complexité de *BS_FD*

La complexité de *BS_FD* est proportionnelle à $l \times m \times n$, en considérant que le générateur de nombres au hasard fonctionne en temps constant. En effet, la complexité de recherche du k^e plus grand élément d'un tableau est proportionnelle à la taille du tableau. La valeur de l doit être assez grande pour que l'imprécision liée à la finitude de l ne nuise pas à la confiance globale, mais elle ne dépend ni de m ni de n. L'algorithme est donc globalement linéaire en $m \times n$, avec une forte constante l, liée au bootstrap.

6 Expérimentations

6.1 Description des données

Les méthodes de filtrage présentées ici ont été appliquées à cinq bases de règles disponibles sur la plateforme HERBS [Vaillant et al. 2003]. Celles-ci ont été extraites à l'aide d'Apriori suivant l'implémentation de [Borgelt et Kruse 2002] de bases de cas du site UCI (http://www.ics.uci.edu/~mlearn/MLSummary.html) : Contraceptive Method Choice (CMC), Flags (Flags), Wisconsin breast Cancer (WBC), Solar Flare I (SFI) et Solar Flare II (SFII). Nous avons calculé pour chaque méthode le taux de réduction de chaque base, après retrait des règles non significatives.

6.2 Caractéristiques et résultats

Le tableau ci-dessous est composé de deux sous-tableaux qui récapitulent les caractéristiques de chaque base et le nombre de règles sélectionnées suivant chaque méthode :

- Contrôle à 5% : on sélectionne les règles ayant une *p-value* $\leq 5\%$.
- Bonferroni : la correction est appliquée sur un seuil de 5%.
- Holm : la procédure est appliquée avec un seuil de 5%.
- *BS_FD* (r) : risque de 5%, avec un V_0 égal au résultat du *pFDR*, appliqué au coefficient de corrélation r comparé à 0, unilatéralement à droite.

Caractéristiques	CMC	Flags	WBC	SF I	SF II
Nb cas	1473	194	699	323	1066
Nb règles	2878	3329	3095	5402	3596
Tx couverture	100%	100%	96.2%	100%	100%
Tx recouvrement	259	1848	646	1828.6	2277
Seuil support	5%	50%	10%	20%	20%
Seuil confiance	60%	90%	70%	85%	85%
Résultats	CMC	Flags	WBC	SF I	SF II
contrôle à 5%	1401	2181	3094	2544	2558
pFDR	916 (3)	1200 (3)	3095 (0)	900 (5)	1625 (4)
FDR (Benj.)	913 (0.003)	1198 (0.0027)	/	899 (0.006)	1626 (0.0022)
BS_FD (r)	794 (3)	1074 (3)	3093 (0)	604 (5)	738 (4)
Holm	742	564	3094	432	1020
Bonferroni	731	539	3042	427	1006

TAB. 2 – *Filtrage de quelques bases de règles*

- *FDR* : la méthode décrite en section 4 est utilisée avec un seuil égal à celui de la dernière *q-value* sélectionnée par le *pFDR*, indiqué entre parenthèses (pour pouvoir comparer grossièrement le *pFDR* et le *FDR*, il est nécessaire de sélectionner les règles à un degré de contrôle voisin).
- *pFDR* : la méthode est utilisée avec une zone de rejet de 0.1%. Entre parenthèses est indiqué le nombre moyen de rejets à tort. La zone de rejet est choisie telle que ce nombre soit le plus acceptable possible.
- BS_FD (m.c) : identique à BS_FD (r) avec trois critères, r (comparé à 0), le support et la confiance (comparés aux seuils utilisés en extraction).

A partir du tableau 2, on opère différentes constatations :

- Sur l'une des bases, WBC, le filtrage est totalement inefficace y compris la correction de Bonferroni. La raison en est qu'une seule règle de la base de départ a une *p-value* supérieure à 0.05 (à savoir 0.184), une autre est à 0.023, les autres *p-value*s sont inférieures à 0.01, 3036 d'entre elles valent 0.0000 !
- Pour les 4 autres bases, le filtrage par la simple répétition du test d'indépendance réduit notablement la base : 51%, 34%, 53%, 39%.
- Dans la base ainsi filtrée, il reste encore beaucoup de faux positifs, que les différentes méthodes de contrôle du risque permettent encore d'éliminer.
- Comme prévu, la procédure de contrôle du risque la plus sévère est la correction de Bonferroni, donnant sur les 4 bases des taux de réduction de 75%, 81%, 65% et 60%. Par sa sévérité même, cette procédure est peu puissante, limitant certes les faux positifs, mais au prix d'une augmentation des faux négatifs. La procédure de Holm donne des résultats voisins, son inefficacité venant du très grand nombre de règles qui rend inutile la correction du seuil pas à pas.
- Les méthodes pFDR, FDR(Benjamini) et notre méthode BS_FD donnent des résultats intermédiaires, correspondant à ce que l'on pouvait attendre. La méthode BS_FD apparaît comme la plus sévère des 3, plus particulièrement sur Solar

Résultats	CMC_app	CMC_val	croisement	Taux
sans contrôle	3391	2742	2742	0.81
contrôle à 5%	1835	1462	1238	0.68
pFDR	1347 (3)	996 (4)	938	0.70
BS_FD (r)	1302 (3)	955 (4)	903	0.69
Holm	1149	858	795	0.69
BS_FD (m.c)	159 (3)	276 (4)	125	0.79

TAB. 3 – *Résultats de la base CMC en validation*

Flare II, mais la raison est que le paramétrage de pFDR et FDR(Benjamini) assure un nombre moyen de fausses découvertes égal à V_0, alors que BS_FD assure que V_0 n'est dépassé qu'avec le risque 0.05, ce qui est plus exigeant.

– La procédure de filtrage est d'autant plus nécessaire qu'elle permet d'éliminer des règles qui seraient sélectionnées avec bon nombre de mesures de qualité. Ainsi les règles logiques dont le conséquent est très fréquent (*e.g.* Solar Flare II) ont-elles une valeur maximale pour toutes les mesures qui donnent une valeur fixe maximale aux règles logiques, alors qu'elles n'ont aucune espèce d'intérêt et que leur *p-value* est non significative. A l'inverse, le calcul des *p-value*s ne préjuge pas du classement ultérieur des règles par des mesures descriptives favorisant les règles les plus intéressantes, par exemple des mesures à la fois dissymétriques et qui avantagent les règles dont le conséquent est rare.

6.3 Validation des résultats

Alors que le découpage de la base de cas en base d'apprentissage et base de validation est courant en apprentissage supervisé, il n'est jamais pratiqué dans le domaine des règles d'association. La raison est sans doute que les règles d'association ressortent de l'apprentissage non supervisé et que leur extraction suivant le support et la confiance est une tâche déterministe clairement définie, comme le souligne [Freitas, 2000]. Dans la mesure où nous avons introduit un test de signification, il est logique de distinguer base d'apprentissage et base de validation pour étudier la pertinence sur la base de test des règles sélectionnées sur la base d'apprentissage.

La base *CMC* a été séparée au hasard pur en 2 bases de même taille. Sur CMC_app, nous avons extrait une base de règles admissibles au sens de l'algorithme *Apriori* et nous avons filtré ces règles à l'aide des différentes méthodes présentées. Ensuite (tableau 3), nous avons examiné comment se comportaient les règles admissibles issues de la base CMC_app (col. 1) lorsqu'elles étaient appliquées à la base CMC_val (col. 2) et nous avons calculé parmi les règles sélectionnées par chaque méthode sur CMC_app combien étaient encore sélectionnées par la même méthode sur CMC_val (col. 3). On constate ainsi que 80% des règles admissibles pour CMC_val le sont encore pour CMC_app, alors que ce taux descend à 70% lorsque l'on filtre la base de règle pour ne garder que les règles significatives, quelle que soit la méthode. En revanche, l'application de BS_FD(m.c) avec les trois critères, support, confiance et r, permet de revenir à ce taux de 80% au prix il est vrai d'une sérieuse diminution du nombre de règles.

7 Conclusion et travaux futurs

Nous partons d'un principe simple, il ne faut garder dans une base de règles que celles qui renforcent la conclusion sur le conséquent. Pour ce faire, nous proposons une stratégie de filtrage des bases de règles qui conduit à pratiquer une multitude de tests unilatéraux à droite sur le coefficient de corrélation entre A et B. Cette stratégie est fondée sur le contrôle du nombre de règles sélectionnées à tort et non pas du risque, assurant par là une plus grande puissance, tout en permettant à l'utilisateur d'arbitrer entre "règles sélectionnées à tort" et "règles pertinentes non sélectionnées". Nous proposons un algorithme original BS_FD qui a l'avantage de contrôler directement le nombre de "faux positifs" et non pas leur moyenne, en tenant compte de la dépendance des règles, tout en permettant de tester plusieurs critères à la fois. Les expérimentations montrent l'efficacité de la stratégie proposée qui permet de réduire sensiblement la taille de la base, facilitant le recours ultérieur à des mesures de qualité descriptives qui apportent un point de vue complémentaire sur la pertinence des règles.

Une extension de travail est prévue en fouille des données génomiques pour rechercher des règles discriminantes. Les procédures de contrôle du risque intéressent l'ensemble des méthodes de fouille des données où l'on multiplie les tests.

Références

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. Fast algorithms for mining association rules. *Proc. of the 20th VLDB Conference*, Santiago, Chile, 1994.

[Agrawal et al., 1993] R. Agrawal, T. Imielinski et A. Swami. Mining associations between sets of items in large databases, *Proc. of the ACM SIGMOD Conf.*, Washington DC, USA, 1993.

[Benjamini et Hochberg, 1995] Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statisc. Soc., B*, 57:289-300, 1995.

[Benjamini et Liu, 1999] Y. Benjamini et W. Liu. A step-down multiple-hypothesis procedure that controls the false discovery rate under independance, J. Stat. Planng Inf., 82:163-170, 1999.

[Borgelt et Kruse, 2002] C. Borgelt et R. Kruse. Induction of association rules : Apriori implementation. *Proc. 15th Conf. on Comp. Stat.*, Physika Verlag, Germany, 2002.

[Efron, 1979] B. Efron. Bootstrap methods: Another look at the jacknkife, *Annals of statistics*, 7:1-26, 1979.

[Freitas, 2000] A. Freitas. Understanding the crucial difference between classification and discovery of association rules. *SIGKDD Explorations*, vol. 2, 1:65-69, 2000.

[Ge et al., 2003] Y. Ge, S. Dudoit et T.P. Speed. Resampling-based multiple testing for microarray data analysis, *Tech. Rep. 663*, Univ. of California, Berkeley, 2003.

[Giné et Zinn, 1984] E. Giné et J. Zinn. Bootstrapping general empirical measures, *Annals of probability*, 18:851-869, 1984.

[Gras, 1979] R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathematiques*, Thèse

d'Etat, Rennes 1, 1979.

[Gras et al., 2001] R. Gras, P. Kuntz, R. Couturier et F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux, *Revue ECA, Extraction des Connaissances et Apprentissage*, Hermès, 1:69-80, 2001.

[Holm, 1979] S. Holm. A simple sequentially rejective multiple test procedure. Scand. J. Statistic., 6:65-70, 1979.

[Lallich et Teytaud, 2003] S. Lallich et O. Teytaud. Evaluation et validation de l'intérêt des règles d'association, à paraître *RNTI, Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse.

[Lenca et al. 2003] P. Lenca, P. Meyer, P. Picouet, B. Vaillant et S. Lallich. Critères d'évaluation des mesures de qualité des règles d'association, *RNTI, Revue des Nouvelles Technologies de l'Information*, 1:123-134, Cépaduès, Toulouse.

[Lerman et Azé, 2003] I.C. Lerman et J. Azé, Une mesure contextuelle dicriminante de qualité des règles d'association, *EGC'03, RIA-ECA*, 17, 1-2-3:247-262, 2002.

[Lerman et al., 1981] I.C. Lerman, R. Gras et H. Rostam. Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines*, 74:5-35, 75:5-47, 1981.

[Storey, 2001] J.D. Storey. The positive false discovery rate and the q-value, écrit en 2001, à paraître *Annals of Statistics* 2003.

[Teytaud et Lallich, 2001] O. Teytaud et S. Lallich. Bornes uniformes en extraction de règles d'association, *Actes Colloque CAp'01*, Grenoble, 133-148, 2001.

[Vaillant et al. 2003] B. Vaillant, P. Picouet et P. Lenca. An extensible platform for rule quality measure benchmarking. R. Bisdorff (Ed), *Human Centered Processes*, 187-191, 2003.

[Van Der Vaart et Wellner, 1996] A. Van Der Vaart et J.A. Wellner. *Weak Convergence and Empirical Processes*, Springer Series in Statistics, 1996.

[Vapnik, 1995] V.N. Vapnik. *The nature of statistical learning*, Springer, 1995.

[Westfall et Young, 1993] P.H. Westfall et S.S. Young. *Resampling based multiple testing : examples and methods for p-values adjustment*, John Wiley & Sons, 1993.

Summary

Association rules extraction algorithms allow to efficiently go through the itemsets lattice in order to constitute a base of rules acceptable at predefined support and confidence levels. However they result in a multitude of rules hardly exploitable. We suggest to refine such bases by eliminating non statistically significant rules. The multitude of performed tests mechanically leads to a multiplication in false discoveries. We first present procedures issued from biostatistic which aim at controlling not the risk, but the number of false discoveries. Then, we propose BS_FD, an original algorithm based on bootstrap, which selects significant rules while controlling the number of false discoveries. By experimenting these procedures on different rule bases, we show their ability to refine rules bases.

Keywords: Association rule, quality, multiple testing risk control.