# Évaluation et validation de l'intérêt des règles d'association

Stéphane Lallich*, 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

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

**Résumé.** La recherche de règles d'association intéressantes est un thème privilégié de l'extraction des connaissances à partir des données. Les algorithmes du type Apriori fondés sur le support et la confiance des règles ont apporté une solution élégante au problème de l'extraction de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes. Il faut disposer d'autre mesures venant compléter le support et la confiance. Dans cet article, nous passons en revue les principales mesures proposées dans la littérature et nous proposons des critères pour les évaluer. Nous suggérons ensuite une méthode de validation qui utilise les outils de la théorie de l'apprentissage statistique, notamment la *VC-dimension*. Face au grand nombre de mesures et à la multitude de règles candidates, l'intérêt de ces outils est de permettre la construction de bornes uniformes non asymptotiques pour toutes les règles et toutes les mesures simultanément.

## 1 Introduction

L'étude des règles d'association entre attributs booléens est déjà ancienne, liée à l'analyse des tableaux croisés $2 \times 2$. Comme le soulignent (Hajek et Rauch 1999), l'une des premières méthodes de recherche des règles d'association est la méthode GUHA initiée par (Hajek, Havel et Chytil 1966), où apparaissent déjà les notions de support et de confiance. L'intérêt pour les règles d'association a été renouvelé par les travaux de (Agrawal, Imielinski et Swami 1993), (Agrawal et Srikant 1994), puis (Srikant et Agrawal 1995) ayant trait à l'extraction de règles d'association à partir des grandes bases de données qui enregistrent le contenu des transactions commerciales.

Dans une telle base, chaque enregistrement est une transaction alors que les différents champs correspondent aux articles susceptibles de composer la transaction. On note $n$ le nombre de transactions et $p$ le nombre d'articles. Dans la mesure où l'on s'intéresse à la présence-absence de chaque article dans les différentes transactions, on

associe à chaque article l'acte d'achat correspondant, appelé item, qui est une variable booléenne. Sur l'ensemble des transactions, on obtient une matrice booléenne de dimensions $n$ et $p$. A un ensemble d'articles, on associe la conjonction des actes d'achat correspondant, ou *itemset*, qui est aussi une variable booléenne.

A partir de la matrice booléenne qui indique les articles présents dans chaque transaction, on veut extraire des règles du type « si un client achète du pain et de la moutarde, alors il est très probable qu'il achète aussi des saucisses ». Une règle d'association est ainsi une expression $r$ du type $A \rightarrow B$, ou l'antécédent $A$ et le conséquent $B$ sont des *itemsets* qui n'ont pas d'items communs.

De façon plus générale, ce formalisme peut s'appliquer à toute base de données dont on a extrait une table individus-variables, à condition de discrétiser les variables continues et de mettre les variables catégorielles sous forme disjonctive complète, ainsi l'exploration du transcriptome en génomique (Becquet et al. 2002). Des recherches plus récentes ont entrepris de travailler directement sur les données non booléennes, ainsi (Guillaume 2000) pour la découverte de règles ordinales.

Dans la mesure où le nombre de règles d'association possibles croît exponentiellement avec le nombre d'items, il est capital de pouvoir se limiter à l'extraction des règles les plus « intéressantes ». Il faut pour cela être capable de définir celles-ci et de les identifier, puis il faut les valider. Nous présentons tout d'abord les algorithmes liés aux critères de support et de confiance (section 2) et nous montrons les limites de cette approche. Nous précisons ensuite la notion de règle par rapport à celle d'implication ou de corrélation (section 3). Après avoir indiqué la définition des principales mesures de l'intérêt des règles (section 4), nous indiquons les critères sur la base desquels on peut évaluer celles-ci (section 5). Nous proposons alors une présentation synthétique de ces mesures (section 6). Enfin, nous abordons le problème de la validation des règles à travers les outils de l'apprentissage statistique (section 7) et nous indiquons quelques perspectives (section 8).

# 2 Intérêt et limites de l'approche support-confiance

Par approche support-confiance, on désigne les algorithmes d'extraction qui recherchent de façon exhaustive les règles d'association dont le support et la confiance dépassent des seuils fixés au préalable par l'utilisateur, notés $min_{supp}$ et $min_{conf}$.

## 2.1 Support et confiance

Soient $n(A)$ 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 $B$. 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 $A$, c'est-à-dire la fréquence relative conditionnelle de $B$ sachant $A$ :

$Conf\,(A \rightarrow B) = \frac{P(AB)}{P(A)} = \frac{n(AB)}{n(A)} = 1 - \frac{n\left(A\overline{B}\right)}{n(A)}$

## 2.2 Algorithmes d'extraction suivant le support et la confiance

Les algorithmes d'extraction liés à l'approche support-confiance parcourent le treillis des *itemsets* pour rechercher les *itemsets* fréquents, ceux dont le support dépasse $min_{supp}$, pour en déduire les règles d'association dont la confiance dépasse $min_{conf}$. En effet, le treillis des *itemsets* admet une double propriété qui rend très efficace la condition de support lors de la recherche des fréquents :

- tout sous-ensemble d'un *itemset* fréquent est fréquent ;
- tout sur-ensemble d'un *itemset* non fréquent est non fréquent.

Apriori, l'algorithme fondateur (Agrawal et Srikant 1994) procède en deux temps :

1. On recherche les *itemsets* fréquents, ceux dont le support dépasse $min_{supp}$, en balayant le treillis des *itemsets* dans sa largeur et en calculant les fréquences par comptage dans la base, ce qui impose une passe sur la base à chaque niveau du treillis ;
2. Pour chaque *itemset* fréquent $X$, on conserve les seules règles du type $X \backslash Y \rightarrow Y$, avec $Y \subset X$, dont la confiance dépasse le seuil $min_{conf}$.

Les règles déduites des *itemsets* fréquents ont nécessairement une confiance supérieure au seuil de support, dans la mesure où $Supp\,(A \rightarrow B) < Conf\,(A \rightarrow B)$.

L'efficacité de Apriori diminue en présence de données denses ou fortement corrélées. Toute la difficulté de l'extraction des fréquents consiste à identifier la bordure entre *itemsets* fréquents et *itemsets* non-fréquents dans le treillis des *itemsets* (Hipp et al. 2000). La recherche peut se faire en largeur dans le treillis ou en profondeur. Dans chaque cas, on peut procéder par comptage direct de la fréquence de chaque *itemset* dans la base, ou procéder par intersection des deux *itemsets* qui constituent l'*itemset* candidat. Parmi les améliorations proposées pour accélérer la construction des ensembles fréquents dans certaines situations, on citera :

- extraction d'un échantillon de la base qui tienne en mémoire, à partir duquel on construit l'ensemble des *itemsets* fréquents dans l'échantillon ainsi que sa bordure négative constituée des *itemsets* non fréquents minimaux dont toutes les parties sont fréquentes (algorithme Sampling, Toivonen 1996), ce qui limite le risque de non exhaustivité ;
- diminution progressive de la base : au lieu de faire une passe lors de l'examen de chaque niveau du treillis des *itemsets*, on met toute la base en mémoire et à chaque niveau du treillis, on représente les transactions par les *k-itemsets* candidats qu'elle contient ; une seule passe suffit donc, mais il faut que toute la base tienne en mémoire (algorithme AprioriTID, Agrawal et Srikant 1994) ;
- dynamisation de l'algorithme : on procède par niveaux dans le treillis, mais au niveau k dès qu'un *itemset* a atteint le seuil de fréquence, on introduit les *itemsets* candidats de niveau $k+1$ qu'il contribue à générer, ce qui diminue le nombre de passes nécessaires sur la base (algorithme DIC, *Dynamic Itemset Counting*, Brin et al. 1997b) ;

- partitionnement de la base tel que les tid-listes (ensemble des tid – ou identifiant unique d'une transaction – associés aux transactions qui contiennent un *itemset* donné) intermédiaires associées au treillis de chaque partie tienne en mémoire ; dans une première passe, on travaille en largeur et on extrait les tid-listes des *itemsets* du niveau $k$ pour construire les *itemsets* fréquents de chaque partie, par intersection des tid-listes du niveau $k-1$ ; dans une seconde passe, on vérifie pour chaque ensemble localement fréquent qu'il est bien globalement fréquent (algorithme Partition, Savasere et al. 1995) ;
- extraction des *itemsets* fermés fréquents, qui constituent une partie génératrice des *itemsets* fréquents et de leur support, ce qui réduit les temps d'extraction et produit des règles non redondantes. Ainsi l'algorithme Close (Pasquier et al. 1999b) et son dérivé A-Close (Pasquier et al. 1999a), puis l'algorithme Pascal fondé sur le comptage par inférence des ensemble clés (Pasquier et al. 2002). Les ensembles fermés et les ensembles clés (ou libres) ont été étudiés aussi par (Boulicaut et Bykowski 2000, Boulicaut et al. 2000) qui ont étendu ces notions à celles d'ensembles $\delta$-fermés et d'ensembles $\delta$-libres.
- recherche en profondeur et comptage dans la base, rendus possibles par une représentation très condensée des données de transaction, appelée FP-tree (algorithme FP-Growth, Han et al. 2000) ;
- recherche en profondeur dans le treillis par intersection rapide des tid-listes, la procédure étant interrompue dès que l'on est sûr que l'*itemset* candidat ne peut plus être fréquent, ainsi Eclat (Zaki et al. 1997).

Nous n'évoquons pas ici les algorithmes de recherche directe d'*itemsets* fréquents maximaux (*itemsets* fréquents dont tous les sur-ensembles sont non-fréquents) dans la mesure où ceux-ci se prêtent mal au calcul du support des ensembles fréquents qu'ils contiennent, calcul pourtant nécessaire à celui de la confiance, ainsi MaxEclat (Zaki et al. 1997) ou Max-Miner (Bayardo 1998).

## 2.3 Avantages et inconvénients des critères de support et de confiance

Par delà leur grand intérêt comme critères d'extraction, on insistera d'abord sur une importante qualité du support et de la confiance qui est leur grande intelligibilité. Le sens concret des valeurs du support et de la confiance est parfaitement assimilable par l'utilisateur non spécialiste.

Selon (Freitas 2000), la tâche des algorithmes d'extraction de règles d'association relevant de l'approche support-confiance se distingue radicalement de celle des algorithmes de classification supervisée. En effet, cette tâche est clairement définie et déterministe, exempte de biais d'induction et de risque de surajustement aux données au sens où tous les algorithmes doivent découvrir les mêmes règles d'association, celles qui vérifient les conditions de support et de confiance préalablement retenues. C'est cette tâche même qui pose question. Si l'approche support-confiance présente un grand intérêt pour l'extraction, son intérêt pour l'utilisateur est plus discutable.

Tout d'abord, les algorithmes liés à cette approche engendrent un très grand nombre de règles qui sont difficiles à gérer et dont beaucoup n'ont que peu d'intérêt ! En outre,

| $A \setminus B$ | 0 | 1 | total |
|---|---|---|---|
| 0 | $P\left(\overline{A}\overline{B}\right)$ | $P\left(\overline{A}B\right)$ | $P\left(\overline{A}\right)$ |
| 1 | $P\left(A\overline{B}\right)$ | $P\left(AB\right)$ | $P\left(A\right)$ |
| total | $P\left(\overline{B}\right)$ | $P\left(B\right)$ | 1 |

TAB. 1 – *Notations pour la distribution des itemsets $A$ et $B$*

| $A\backslash B$ | 0 | 1 | total |
|---|---|---|---|
| 0 | $1 - c/l - s/c + s$ | $c/l - s$ | $1 - s/c$ |
| 1 | $s/c - s$ | $s$ | $s/c$ |
| total | $1 - c/l$ | $c/l$ | 1 |

TAB. 2 – *Distribution de $A$ et $B$ en fonction du support $s$, de la confiance $c$ et du lift $l$*

la condition de support qui est le moteur même du processus d'extraction écarte les règles ayant un petit support alors que certaines peuvent avoir une très forte confiance et présenter un réel intérêt, le cas est courant en marketing (les pépites du *Data Mining*). Si l'on baisse le seuil de support pour remédier à cet inconvénient, les ensembles fréquents sont trop nombreux et les algorithmes d'extraction sont asphyxiés.

Enfin, les seules conditions de support et de confiance ne suffisent pas à assurer le réel intérêt d'une règle. En effet, une règle $A \rightarrow B$ dont la confiance est égale à la probabilité de $B$, soit $P(B/A) = P(B)$ ce qui est la définition de l'indépendance de A et B, n'apporte aucune information ! Par exemple, si $P(A) = 80\%$ et $P(B) = 90\%$, la règle $A \rightarrow B$ a un support égal à 72 % et une confiance de 90 % en cas d'indépendance.

En résumé, il faut au minimum prendre en compte d'autres mesures d'intérêt des règles que le support et la confiance, favorisant ainsi un biais d'induction, d'où l'importance de réfléchir à la nature particulière des règles d'association si l'on veut établir un banc d'essai des principales mesures d'intérêt.

## 3 Règle, implication et équivalence

Pour bien comprendre la différence entre les notions de règle d'association, d'implication et de corrélation, prenons le cas de deux *itemsets* $A$ et $B$ dont le tableau conjoint est donné par le tableau 1, où par abus de langage 0 signifie *faux* et 1 signifie *vrai*.

En premier lieu, on soulignera qu'un tel tableau n'a que 3 degrés de liberté lorsque ses marges ne sont pas fixées, au sens où la connaissance de 3 probabilités suffit à reconstruire le tableau. C'est ainsi que si l'on considère trois mesures non liées, par exemple le support $s$, la confiance $c$ et le lift $l$ (Brin et al. 1997a), défini par $l(A \rightarrow B) = \frac{P(AB)}{P(A)P(B)}$, la connaissance de $s$, $c$ et $l$, où $s < c < l$, définit entièrement la distribution de probabilités conjointes de $A$ et $B$ (tableau 2).

Dans la logique des promoteurs des règles d'association, exprimée par l'approche

| $A \to B$ | | $\overline{B} \to \overline{A}$ | | $B \to A$ | | $\overline{A} \to \overline{B}$ | | $A \to \overline{B}$ | | $B \to \overline{A}$ | | $\overline{B} \to A$ | | $\overline{A} \to B$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| o | o | + | o | o | - | + | - | o | o | o | + | - | o | - | + |
| - | + | - | o | o | + | o | o | + | - | o | - | + | o | o | o |

| $A \Longrightarrow B$ | | $B \Longrightarrow A$ | | $A \Longrightarrow \overline{B}$ | | $\overline{B} \Longrightarrow A$ | |
|---|---|---|---|---|---|---|---|
| $\overline{B} \Longrightarrow \overline{A}$ | | $\overline{A} \Longrightarrow \overline{B}$ | | $B \Longrightarrow \overline{A}$ | | $\overline{A} \Longrightarrow B$ | |
| + | + | + | - | + | + | - | + |
| - | + | + | + | + | - | + | + |

| $A \Leftrightarrow B$ | | $A \Leftrightarrow \overline{B}$ | |
|---|---|---|---|
| $\overline{B} \Leftrightarrow \overline{A}$ | | $B \Leftrightarrow \overline{A}$ | |
| + | - | - | + |
| - | + | + | - |

TAB. 3 – *Exemples et contre-exemples des règles, implications et équivalences*

support-confiance, on ne s'intéresse qu'aux exemples de $A$, à leur fréquence globale (support) et à leur répartition entre $B$ et $\overline{B}$ (confiance). La répartition des exemples de $\overline{A}$ entre $B$ et $\overline{B}$ n'est pas prise en compte. Le premier bloc de lignes du tableau 3 (lignes 1, 2 et 3) indique pour chacune des 8 règles que l'on peut considérer à partir du couple ($A$,$B$) la nature – exemples (+), contre-exemples (-), ou cas non pris en compte (o) – de l'état (0 ou 1) du conséquent (colonnes) suivant l'état (0 ou 1) de l'antécédent (lignes). On constate ainsi qu'une règle et sa contraposée ont les mêmes contre-exemples, mais qu'elles diffèrent par les exemples. On pourra considérer que les 4 premières règles du bloc 1 sont covariantes, alors que les 4 suivantes sont contravariantes. Le second bloc de lignes (lignes 4, 5, 6 et 7) du tableau 3 rassemble les mêmes informations pour les 4 implications logiques, alors que le troisième bloc (lignes 8, 9, 10, 11) concerne les deux équivalences.

L'implication logique $A \Longrightarrow B$ et sa contraposée $\overline{B} \Longrightarrow \overline{A}$ correspondent à $\overline{A} \vee B$, avec $A\overline{B}$ comme seule situation correspondant à des contre-exemples, les autres correspondant à des exemples. Enfin, l'équivalence logique $A \Leftrightarrow B$ et sa contraposée $\overline{B} \Leftrightarrow \overline{A}$ correspondent à $(AB) \vee \left(\overline{AB}\right)$, avec pour exemples (resp. contre-exemples) les exemples (resp. contre-exemples) des 4 règles covariantes. On trouve dans (Kodratoff, 1999) un intéressant développement sur la distinction entre dépendance descriptive et dépendance causale. Il faut noter que la notion de règle exposée ici dans la logique du couple support-confiance ne suffit pas à définir le réel intérêt d'une règle, puisque la confiance de la règle n'a de sens que comparée à $P(B)$, dans le contexte global.

# 4 Liste des mesures

Nous rassemblons dans le tableau 4 les principales mesures de l'intérêt d'une règle auxquelles nous nous référons dans la suite de ce travail.

| Nom | Formule | Réf. |
|---|---|---|
| Confiance | $P(B/A)$ | Agr 93 |
| Confiance centrée | $P(B/A) - P(B)$ | |
| Pearl | $P(A)\,\lvert P(B/A) - P(B)\rvert$ | Pea 88 |
| Piatetsky-Shapiro | $nP(A)\,(P(B/A) - P(B))$ | Pia 91 |
| Loevinger : | $\frac{P(B/A)-P(B)}{P(\overline{B})}$ | Loe 47 |
| Zhang : | $\frac{P(AB)-P(A)P(B)}{Max\left\{P(AB)P(\overline{B});P(B)(P(A)-P(AB)\right\}}$ | Zha 00 |
| Corrélation | $\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$ | |
| Indice d'implicat. | $\sqrt{n}\frac{P(A\overline{B})-P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$ | Ler 81 |
| Lift | $\frac{P(AB)}{P(A)P(B)}$ | Bri 97a |
| Surprise | $\frac{P(AB)-P(A\overline{B})}{P(B)}$ | Aze 02 |
| Conviction | $\frac{P(A)P(\overline{B})}{P(A\overline{B})}$ | Bri 97a |
| Intensité d'implicat. | $P\left[Poisson\left(nP\left(A\right)P\left(\overline{B}\right)\right) \geq nP(A\overline{B})\right]$ | Gra 79 |
| Sebag-Schoenauer | $\frac{P(AB)}{P(A\overline{B})}$ | Seb 88 |
| Multiplicateur de cote | $\frac{P(AB)P(\overline{B})}{P(A\overline{B})P(B)}$ | Lal 02 |
| *J-mesure* | $P(AB)\log\frac{P(AB)}{P(A)P(B)} + P(A\overline{B})\log\frac{P(A\overline{B})}{P(A)P(\overline{B})}$ | Goo 88 |

TAB. 4 – *Présentation des principales mesures d'intérêt*

# 5 Critères d'appréciation

Nous avons montré la nécessité de compléter le support et la confiance par d'autres mesures d'intérêt, qui seront principalement celles du tableau 4, pour mettre en évidence le caractère inattendu d'une règle compte tenu des fréquences respectives de $A$ et $B$. Nous allons évoquer les critères selon lesquels on peut apprécier de telles mesures $m(A \rightarrow B)$, ce qui nous permettra dans la section suivante une analyse critique des mesures d'intérêt usuelles. Une démarche comparable a été menée par (Tan et al. 2002) mais elle s'adresse à des mesures symétriques ou symétrisées, alors que nous nous intéressons en priorité aux mesures non symétriques (cf. section 5.2).

## 5.1 Sens concret de la mesure

La mesure a-t-elle un sens concret qui soit parlant pour l'utilisateur ? C'est bien sûr le cas du support et de la confiance, mais aussi celui du lift, de la conviction, de la mesure de Sebag et Schoenauer ou de la mesure *mc*. Un lift égal à 2 signifie que le nombre d'exemples de la règle $A \rightarrow B$ est 2 fois plus grand que celui attendu sous l'indépendance, ce qui implique que le consommateur qui achète $A$ a 2 fois plus de chances d'acheter $B$ que le consommateur en général, mais aussi l'inverse puisque le lift est symétrique et que les exemples de $A \rightarrow B$ sont aussi ceux de $B \rightarrow A$. Une conviction égale à 2 signifie que le nombre de contre-exemples de la règle (situation $A\overline{B}$)

est deux fois moins grand que celui attendu sous l'indépendance de $A$ et $B$. Lorsque la mesure de Sebag et Schoenauer vaut 2, la cote de l'achat de $B$ en cas d'achat de $A$ vaut 2, ce qui signifie qu'un consommateur qui achète $A$ a 2 fois plus de chance d'acheter $B$ que de ne pas acheter $B$, soit 2 chances sur 3 d'acheter $B$. Une valeur égale à 2 de la mesure *mc* signifie que la cote de l'achat de $B$ est multipliée par 2 en cas d'achat de $A$. L'interprétation des autres mesures est moins aisée, notamment la *J-mesure* et l'intensité d'implication, plus particulièrement sous sa forme entropique.

## 5.2 Mesure et règle visée

Une mesure doit distinguer les différentes règles associant $A$ et $B$ (tableau 3).

1. La mesure doit impérativement permette de choisir entre $A \rightarrow B$ et $A \rightarrow \overline{B}$, les exemples de l'une étant les contre-exemples de l'autre, ce que ne font pas le $\chi^2$, la mesure de Pearl ou la *J-mesure* au contraire du coefficient de corrélation $r$.
2. On préférera les mesures dissymétriques qui respectent la nature des règles d'association transactionnelles : "si tels articles (A) sont dans le panier, alors le plus souvent tels autres (B) y sont". Les mesures symétriques comme le support, la mesure de Piatetsky-Shapiro, le lift ou $r$ et ses dérivés, évaluent de la même façon les règles $A \rightarrow B$ et $B \rightarrow A$, alors que celles-ci ont les mêmes exemples mais pas les mêmes contre-exemples.
3. Une mesure doit-elle évaluer de la même façon $A \rightarrow B$ et $\overline{B} \rightarrow \overline{A}$ (Kodratoff 1999) ? Si l'égalité stricte est requise au sens de l'implication logique, elle ne l'est pas au sens des règles d'association. En effet, les deux règles ont les mêmes contre-exemples, mais elles n'ont pas les mêmes exemples. La prise en compte de la contraposée, ainsi dans l'intensité d'implication entropique (Gras et al. 2001), rapproche la règle de l'implication logique.

## 5.3 Exemples et contre-exemples

A priori, on peut juger qu'une règle est inattendue en s'intéressant aussi bien au caractère exceptionnel du nombre d'exemples confirmant la règle qu'à celui du nombre d'exemples qui la contredisent. Cependant, on observe (tableau 3) que les exemples de $A \rightarrow B$ sont aussi ceux de $B \rightarrow A$, alors que les contre-exemples de $A \rightarrow B$ sont aussi ceux de $\overline{B} \rightarrow \overline{A}$, ce qui justifie une préférence pour les contre-exemples. Pour que la différence entre les deux points de vue soit réelle, il faut associer à l'étude des contre-exemples une modélisation qui ne fixe pas $n(A)$ car les nombres d'exemples et de contre-exemples sont liés quand la marge de $A$ est fixée : $n(AB) + n(A\overline{B}) = n(A)$.

## 5.4 Sens de variation de la mesure et valeurs de référence

Moins une règle a de contre-exemples, plus elle est intéressante. Une mesure $m$ doit donc être décroissante en fonction du nombre de contre-exemples à marges fixées. Une règle est d'autant plus intéressante que son nombre de contre-exemples (resp. exemples) est exceptionnellement bas (resp. haut) sous l'hypothèse d'indépendance de $A$ et $B$. La

| **Situation** | **Incompatib.** | **Indépend.** | **Règle logique** |
|---|---|---|---|
| Caractérisation | $AB = \emptyset$ | $\frac{P(AB)}{P(A)P(B)} = 1$ | $A \subset B$ |
| Confiance | 0 | $P(B)$ | 1 |
| Confiance centrée | $-P(B)$ | 0 | $P(\overline{B})$ |
| Pearl | $P(A)P(B)$ | 0 | $P(A)P(\overline{B})$ |
| Piatetsky-Shapiro | $-nP(A)P(B)$ | 0 | $nP(A)P(\overline{B})$ |
| Loevinger | $\frac{-P(B)}{P(\overline{B})}$ | 0 | 1 |
| Zhang | $-1$ | 0 | 1 |
| Corrélation | $-\sqrt{\frac{P(A)P(B)}{P(\overline{A})P(\overline{B})}}$ | 0 | $\sqrt{\frac{P(A)P(\overline{B})}{P(\overline{A})P(B)}}$ |
| Indice d'implication (-) | $-P(B)\sqrt{\frac{nP(A)}{P(\overline{B})}}$ | 0 | $\sqrt{nP(A)P(\overline{B})}$ |
| Lift | 0 | 1 | $\frac{1}{P(B)}$ |
| Surprise | $-\frac{P(A)}{P(B)}$ | $2P(A) - \frac{P(A)}{P(B)}$ | $\frac{P(A)}{P(B)}$ |
| Conviction | $P(\overline{B})$ | 1 | $\infty$ |
| Intensité d'implication | 0 | 0.5 | 1 |
| Sebag-Schoenauer | 0 | $\frac{P(B)}{1-P(B)}$ | $\infty$ |
| Multiplicateur de cote | 0 | 1 | $\infty$ |
| *J-mesure* | $P(A)\log\left(\frac{1}{P(\overline{B})}\right)$ | 0 | $P(A)\log\left(\frac{1}{P(B)}\right)$ |

TAB. 5 – *Comportement des principales mesures dans les situations remarquables*

mesure $m$ doit avoir une valeur d'autant plus grande que la règle est plus inattendue par rapport à l'indépendance, dans le sens de la confirmation.

La valeur de la mesure en cas d'indépendance joue le rôle de valeur de référence. Selon (Piatetsky-Shapiro 1991), une bonne mesure doit être :

a. $= 0$, en cas d'indépendance de $A$ et $B$
b. $> 0$, en cas d'attraction, $P(AB) > P(A)P(B)$
c. $< 0$, en cas de répulsion, $P(AB) < P(A)P(B)$

A cet effet, il propose une mesure qui a l'inconvénient d'être symétrique et de ne pas varier entre des bornes fixes, tout en faisant intervenir le nombre de transactions :
$PS(A \to B) = nP(A)\left[P(B/A) - P(B)\right] = n\left[P(AB) - P(A)P(B)\right]$

On constate que pour toutes les mesures présentées dans le tableau 5 la valeur de référence pour l'indépendance est 0 ou 1, sauf pour la confiance et pour la mesure de Sebag-Schoenauer, définie par $Seb(A \to B) = \frac{P(AB)}{P(A\overline{B})}$. On améliore cette dernière en formant (Lallich 2002) la mesure $mc$ (multiplicateur de cote) :

$$mc(A \to B) = \frac{P(\overline{B})}{P(B)} Seb(A \to B) = \frac{P(AB)P(\overline{B})}{P(A\overline{B})P(B)} = l(A \to B) \times conv(A \to B)$$

On peut remplacer b. et c. par des conditions de normalisation b'. et c'. (Zhang 2000) :

b'. $= 1$, au cas où la règle est logique (confiance égale à 1, soit $A \subset B$)

c'. $= -1$, en cas d'incompatibilité (confiance nulle, soit $AB = \emptyset$)

et construit en conséquence la mesure :
$z(A \rightarrow B) = \frac{P(A/B)-P(A/\overline{B})}{Max\{P(A/B);P(A/\overline{B})\}} = \frac{P(AB)-P(A)P(B)}{Max\{P(AB)(1-P(B));P(B)(P(A)-P(AB)\}}$

Les seules mesures présentant des valeurs de référence fixes pour l'indépendance et les valeurs extrêmes (tableau 5) sont la mesure de Zhang (Zhang 2000) et la mesure *mc*. On notera cependant que la valeur en cas d'incompatibilité a beaucoup moins d'importance que celle affectée à la situation de règle logique.

## 5.5 Variation non linéaire

Pour certains auteurs (Gras et al. 2001), il est souhaitable que la mesure $m$ varie de façon non linéaire en fonction de l'apparition des exemples ou des contre-exemples, variation lente au début pour tenir compte du bruit, plus rapide ensuite et à nouveau moins rapide (concavité puis convexité).

Ce n'est pas le cas de la confiance et de toutes les mesures qui s'en déduisent par transformée affine ne dépendant que des marges (tableau 7), puisque la confiance est une fonction affine du nombre d'exemples ou du nombre de contre-exemples à marge de $A$ fixée : $Conf\,(A \rightarrow B) = \frac{n(AB)}{n(A)} = 1 - \frac{n\left(A\overline{B}\right)}{n(A)}$.

A l'inverse, si l'on veut pénaliser les faux positifs, on choisit une mesure comme *mc* qui décroît rapidement en cas d'apparition de contre-exemples (convexe pour les valeurs de $n\left(A\overline{B}\right)$ au voisinage de 0).

## 5.6 Impact de la rareté du conséquent

Une mesure $m$ doit être une fonction croissante de $1-P(B)$ la rareté du conséquent. En effet, plus le conséquent $B$ est rare, plus le fait qu'il contienne l'antécédent $A$ a de l'intérêt. Ceci est particulièrement vrai lorsque l'on s'affranchit de la condition de support. C'est ce qui est fait en partie lorsque l'on utilise une mesure qui résulte d'un centrage de la confiance sur $P(B)$. Cela est aussi obtenu en divisant par $P(B)$ ou en multipliant par $1 - P(B)$, ainsi la mesure *mc* définie par $mc(A \rightarrow B) = \frac{P(\overline{B})}{P(B)} Seb(A \rightarrow B)$ améliore-t-elle la mesure *Seb* sur ce critère.

## 5.7 Approche descriptive *vs.* statistique

On appelle mesure descriptive une mesure qui ne change pas en cas de dilatation des données, lorsque tous les effectifs sont multipliés par un même facteur $\theta$, $\theta > 1$. Sinon la mesure est dite statistique. En première instance, il est logique de préférer une approche statistique, un résultat n'ayant pas la même signification suivant le nombre d'observations $n$ dont il est issu. Une mesure statistique suppose que l'on ait un modèle aléatoire et une hypothèse $H_0$ exprimant l'indépendance de $A$ et $B$. On peut considérer que la base est un modèle de l'échantillon ou que l'aléa est dans la répartition des 0 et des 1 observés pour chaque attribut.

Dans une approche statistique, on choisit une grandeur brute, ainsi le nombre de contre-exemples (ou le nombre d'exemples), puis on en calcule la moyenne et la variance

sous l'hypothèse d'absence de lien. Cette grandeur est ensuite centrée-réduite sous $H_0$, ce qui peut modifier le classement des règles. On est ainsi ramené à la réalisation d'une variable approximativement normale centrée-réduite sous $H_0$, si $n$ assez grand. La fonction de répartition de la loi $N(0,1)$ permet éventuellement de se ramener à un indice entre 0 et 1 en prenant comme mesure la probabilité cumulée ou rétro-cumulée de la variable centrée-réduite observée, ce qui constitue une simple anamorphose monotone qui ramène la mesure à une échelle entre 0 et 1, uniforme sous $H_0$. Lorsque la mesure se prête à une approche statistique où l'aléa réside dans le choix des transactions ou dans une perturbation aléatoire, on peut envisager de calculer un intervalle de confiance pour la mesure théorique à partir de la mesure empirique.

La construction d'un indice d'implication par normalisation de la quantité $n(AB)$ sous l'hypothèse d'absence de lien est proposée par (Lerman, Gras et Rostam 1981). Suivant la modélisation retenue pour exprimer cette hypothèse, à 1, 2 ou 3 niveaux d'aléa, la loi de référence est une loi hypergéométrique $H(n,nP(A),P(B))$, une binomiale $B(n,P(A)P(B))$ ou une loi de Poisson $P(nP(A)P(B))$. Lors de la normalisation, la moyenne est la même, seule change la variance, c'est-à-dire le dénominateur de l'indice normalisé. On obtient 3 indices normalisés, $q_i(A,B), i = 1,2,3$.

La forme 1 se ramène au coefficient de corrélation de points et à la statistique du khi 2, notée $\chi^2$ :

$q_1(A,B) = \sqrt{n}\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$

d'où $q_1^2(A,B) = nr^2(A,B) = \chi^2$. L'indice ainsi obtenu est symétrique et ne change pas lorsque l'on remplace $A$ par $\overline{A}$ et $B$ par $\overline{B}$.

La forme 3, associée à la loi de Poisson, est plus intéressante que la forme 2, car elle traite $(A,B)$ et $(\overline{A},\overline{B})$ de la façon la plus dissymétrique :

$q_3(A,B) = \sqrt{n}\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(B)}}$

Comme le soulignent (Lerman et al. 1981), $q_3(A,B)$ correspond à la contribution orientée de la cellule $AB$ au $\chi^2$ global .

Il est ainsi naturel d'utiliser comme mesure l'opposé de l'indice d'implication associé à $A\overline{B}$ :

$impl(A \rightarrow B) = q_3(A,\overline{B}) = \sqrt{n}\frac{P(A\overline{B})-P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$

Plus récemment, (Lerman et Azé 2002) ont proposé de prendre comme indice d'implication hors contexte la forme descriptive $-\frac{q_3(A,\overline{B})}{\sqrt{n}}$. Celle-ci est l'opposé de la contribution orientée de la cellule $A\overline{B}$ à $\frac{\chi^2}{n}$ et elle varie entre -1 et 1.

L'intensité d'implication, proposée par (Gras 1979) puis (Gras et Lahrer 1993), est fondée sur l'étonnement statistique que provoque le nombre de contre-exemples de la règle, à savoir l'effectif de $A\overline{B}$, sous l'hypothèse d'indépendance de $A$ et $B$, notée $H_0$. Elle correspond ainsi à la probabilité rétrocumulée de la valeur observée de $n(A\overline{B})$, calculée d'après la loi de Poisson $Poisson(nP(A)P(\overline{B}))$, suivant la modélisation à 3 niveaux d'aléa déjà évoquée.

On peut aussi retenir la présentation qui suit. Sous l'hypothèse nulle, tout se passe comme si les 1 de la colonne $A$ et les 1 de de la colonne $B$ avaient été attribués au hasard et indépendamment entre les deux colonnes. Dans le cas du *tirage sans remise*, on exige que le nombre de 1 attribué soit exactement celui observé pour chaque

| (a) | (a) | (a),(b),(c) | (a),(b),(c) | (a) | (b) | (c) | (a) | (b) | (c) |
|---|---|---|---|---|---|---|---|---|---|
| nb ex | nb c-ex | $conf.$ | $r$ | $r^{cr}$ | $r^{cr}$ | $r^{cr}$ | $m$ | $m$ | $m$ |
| 0 | 6 | 0 | -0.707 | -3 | -4.2 | -9.5 | 0,001 | 0,000 | 0.000 |
| 1 | 5 | 0.167 | -0.471 | -2 | -2.8 | -6.3 | 0,023 | 0,002 | 0.000 |
| 2 | 4 | 0.333 | -0.236 | -1 | -1.4 | -3.2 | 0,159 | 0,079 | 0.001 |
| 3 | 3 | 0.5 | 0 | 0 | 0 | 0 | 0,500 | 0,500 | 0.5 |
| 4 | 2 | 0.667 | 0.236 | 1 | 1.4 | 3.2 | 0,841 | 0,921 | 0.999 |
| 5 | 1 | 0.833 | 0.471 | 2 | 2.8 | 6.3 | 0,977 | 0,998 | 1.000 |
| 6 | 0 | 1 | 0.707 | 3 | 4.2 | 9.5 | 0,999 | 1,000 | 1.000 |

TAB. 6 – *Illustration de la dilatation sur un exemple*

variable $A$ et $B$, soit $n(A)$ et $n(B)$, alors que dans le cas du *tirage avec remise*, on décide d'attribuer 0 ou 1 à une transaction avec la probabilité $P(A)$ pour $A$ et $P(B)$ pour $B$. Sous $H_0$, le nombre de contre-exemples suit une loi hypergéométrique (*tirage sans remise*) ou binomiale (*tirage avec remise*). On préférera le *tirage avec remise* qui assure une mesure non symétrique. Le nombre $K$ de contre-exemples de la règle suit alors la loi binomiale $Bin(n,P(A)P(\overline{B}))$ que l'on peut approximer par une loi de Poisson ou par une loi normale, suivant le cas. Le caractère exceptionnel du nombre de contre-exemples observés est caractérisé par la *p-value* unilatérale à gauche d'une loi binomiale de paramètres $n$ et $P(A)P(\overline{B})$ en $nP(A\overline{B})$ et l'intensité d'implication $\varphi(A,B)$ est alors le complément à 1 de cette *p-value*, soit :

$$\varphi(A \rightarrow B) = P\left[Bin(n,P(A)P(\overline{B})) \geq nP(A\overline{B})\right] = P\left[Poi(nP(A)P(\overline{B})) \geq nP(A\overline{B})\right]$$

La difficulté est que pour $n$ assez grand, tout écart à l'indépendance aussi minime qu'il soit est très significatif, ce qui rend tout test illusoire et donne une avalanche de règles difficiles à discerner.

## 5.8 Pouvoir discriminant

Les mesures issues d'une approche statistique ont tendance à perdre leur pouvoir discriminant lorsque le nombre de transactions $n$ est grand. Considérons l'exemple du tableau 6 où les marges sont fixées, $P(A) = \frac{1}{3}$ et $P(B) = \frac{1}{2}$. On part d'un effectif $n = 18$ (situation (a)), et on étudie en faisant varier le nombre de contre-exemples, le comportement de la confiance $conf(A \rightarrow B)$, du coefficient de corrélation $r$, de $r^{CR}$ ce même coefficient centré-réduit sous l'hypothèse d'indépendance, et de la mesure statistique $m$ mesurant le caractère exceptionnel de $r$ en référence à son approximation normale sous l'hypothèse d'indépendance. On recommence les calculs en multipliant les effectifs par 2 (situation (b)), puis par 10 (situation (c)). Il apparaît clairement que plus $n$ augmente, moins la mesure statistique $m$ permet de discerner l'intérêt des règles possibles. En revanche, le classement reste le même. Face à cette difficulté, puisque $n$ est le même pour toutes les règles d'une même base, on peut préférer sélectionner d'abord les règles qui amènent à refuser l'indépendance en direction d'une dépendance positive puis considérer des mesures descriptives centrées et raisonner sur le classement induit par ces mesures.

L'approche contextuelle, développée par Lerman en classification par la vraisem-

blance du lien apporte une première solution au problème de la perte de discrimination des mesures statistiques, comme en témoigne l'indice probabiliste discriminant (Lerman et Azé 2002). Cet indice résulte d'un calcul de probabilité critique après centrage-réduction de l'indice d'implication par rapport à une base de règles admissibles. Cette base $\mathcal{B}$ peut contenir toutes les règles ou de façon plus sélective se limiter à celles qui vérifient certaines conditions, par exemple les conditions de support et de confiance, voire la condition supplémentaire $n(A) < n(B)$. Si l'on note $\Phi$ la fonction de répartition de la loi normale centrée réduite, l'indice probabiliste discriminant $IPD$ est défini par :

$$IPD(A \to B) = 1 - \Phi\left[impl(A \to B)^{CR/\mathcal{B}}\right].$$

Une autre solution consiste à corriger cette perte de pouvoir discriminant. Dans le cas de l'intensité d'implication $\varphi$, Gras, Kuntz, Couturier et Guillet (2001) suggèrent la prise en compte d'un indice d'inclusion, fonction de l'entropie des expériences $B/A$ et $\overline{A}/\overline{B}$. Ils définissent l'indice d'inclusion par $i(A \subset B) = \left[(1 - H(B/A))\left(1 - H(\overline{A}/\overline{B})\right)\right]^{\frac{1}{2}}$ où la fonction $H(B/A)$ vaut :

$= 1 + \frac{1}{2}\left[P(B/A)\log_2 P(B/A) + P(\overline{B}/A)\log_2 P(\overline{B}/A)\right]$, si $P(B/A) < 0.5$,

$= -\frac{1}{2}\left[P(B/A)\log_2 P(B/A) + P(\overline{B}/A)\log_2 P(\overline{B}/A)\right]$, si $P(B/A) \geq 0.5$.

L'intensité entropique s'écrit alors :

$\psi(A \to B) = [\varphi(A \to B) \times i(A \subset B)]^{\frac{1}{2}}$

## 5.9 Fixation d'un seuil

Il est important que les mesures utilisées se prêtent à la fixation d'un seuil qui permette de ne conserver que les règles intéressantes, sans qu'il soit besoin de toutes les classer. Classiquement, on définit ce seuil en référence à la probabilité cumulée de la valeur observée de la mesure sous $H_0$ pour une modélisation donnée. On notera que ce seuil n'est pas un risque compte tenu de la multitude de tests effectués, mais seulement un paramètre de contrôle (Lallich 2002). De par leur définition, la mesure IPD et l'intensité d'implication permettent de fixer directement un tel seuil. Pour les autres mesures, le calcul d'un seuil est un peu moins immédiat. Il devient complexe pour la mesure de Zhang en raison de sa normalisation et pour l'intensité d'implication entropique, en raison de son facteur correctif.

## 5.10 Classement induit par une mesure

Deux mesures $m$ et $m\prime$ classent dans le même ordre les règles extraites d'une base de données si et seulement si pour tout couple de règles extraites de la base :

$m(A \to B) > m(A' \to B') \Longleftrightarrow m'(A \to B) > m'(A' \to B')$

On définit ainsi une relation d'équivalence sur l'ensemble des mesures possibles. Par exemple, la mesure de Sebag-Schoenauer classe comme la confiance, puisqu'elle s'écrit comme une transformation monotone croissante de la confiance : $Seb(A \to B) = \frac{Conf(A \to B)}{1 - Conf(A \to B)}$. De la même façon, on montre que la mesure de Loevinger classe comme la conviction et que la mesure IPD classe comme l'indice d'implication. On ajoutera que si le conséquent est donné, soit $P(B)$ fixé, ce qui est le cas lorsque l'on utilise les

règles d'association en apprentissage supervisé, alors le lift, la conviction, la mesure de Loevinger et la mesure de Sebag-Schoenauer classent comme la confiance !

# 6 Différentes mesures de l'intérêt d'une règle

Dans l'optique support-confiance, le rôle du support est avant tout de rendre praticable l'algorithme d'extraction, du fait de la propriété d'antimonotonie du treillis des fréquents. C'est la confiance qui permet de sélectionner les règles intéressantes parmi celles qui satisfont à la condition de support. Nous avons vu que l'usage d'une autre mesure au moins est nécessaire pour repérer les règles intéressantes. Nous allons préciser la définition des mesures non encore évoquées, en distinguant celles qui sont des fonctions affines de la confiance et les autres.

Après avoir remarqué que le support n'est rien d'autre que l'indice de Russel et Rao (1940), on notera que les indices de proximité usuels définis sur les variables logiques (Lerman 1988) ne sont pas très pertinents dans le cas de l'évaluation des règles d'association en raison de la façon symétrique dont ils traitent les attributs booléens, ainsi l'indice de Ochiai (1957) ($\frac{P(AB)}{\sqrt{P(A)P(B)}}$), celui de Czekanowski-Dice (1913) ($\frac{2P(AB)}{P(AB)+1-P(\overline{AB})}$), ou celui de Kulczynski (1928) ($\frac{P(AB)}{P(A\overline{B})+P(\overline{A}B)}$).

## 6.1 Transformées affines de la confiance

Nombreuses sont les mesures d'intérêt qui s'écrivent comme une normalisation de la confiance par le biais d'une transformation affine $m = \theta_1 (c - \theta_0)$, dont les paramètres ne dépendent que des marges relatives de la table qui croise $A$ et $B$ et éventuellement de $n$ (tableau 7). Le plus souvent, tout revient à un centrage-réduction, où le changement d'origine amène à comparer $P(B/A)$ à $P(B)$, sa valeur attendue en l'absence de liaison, alors que le changement d'échelle varie suivant le but poursuivi. A l'inverse, la lecture des changements d'échelle permet de savoir ce qui différencie deux mesures centrées sur $P(B)$. Deux exceptions, le lift dans lequel la comparaison à $P(B)$ se fait par un simple changement d'échelle et la surprise qui centre la confiance sur 0.5.

Toutes ces mesures corrigent la principale critique faite à la confiance, mais elles héritent par construction de différentes caractéristiques de la confiance. A marges fixées, elles sont une fonction affine du nombre de contre-exemples. Par ailleurs, lorsque le facteur d'échelle ne dépend pas de $n$ (ce qui est le cas de toutes les mesures présentées en 5.2, sauf Piatetsky-Shapiro et l'indice d'implication), ces mesures sont invariantes en cas de dilatation des données. On pourra se reporter à (Guillaume, 2000) pour une étude détaillée, mesure par mesure.

Le lift (Brin et al. 1997a), défini par $l(A \rightarrow B) = \frac{P(AB)}{P(A)P(B)}$, s'interprète comme le quotient du nombre d'exemples observé par celui attendu sous l'hypothèse d'indépendance de $A$ et $B$. S'exprimant à partir des seuls exemples, il est symétrique, puisque les règles $A \rightarrow B$ et $B \rightarrow A$ ont les mêmes exemples.

La surprise (Azé et Kodratoff 2002) est encore une transformée affine de la confiance, mais centrée sur 0.5 et non pas sur $P(B)$, ce qui favorise la prédiction par rapport au ciblage.

| **Mesure** | **centr. ($\theta_0$)** | **éch. ($\theta_1$)** |
|---|---|---|
| Confiance centrée | $P(B)$ | 1 |
| Pearl | $P(B)$ | $P(A)$ |
| Piatetsky-Shapiro | $P(B)$ | $nP(A)$ |
| Loevinger | $P(B)$ | $\frac{1}{P(\overline{B})}$ |
| Zhang | $P(B)$ | $\frac{P(A)}{Max}$ |
| Corrélation | $P(B)$ | $\frac{\sqrt{P(A)}}{\sqrt{P(\overline{A})P(B)P(\overline{B})}}$ |
| Indice implication (-) | $P(B)$ | $\sqrt{n}\sqrt{\frac{P(A)}{P(\overline{B})}}$ |
| Lift | 0 | $\frac{1}{P(B)}$ |
| Surprise | 0.5 | $2^{\frac{P(A)}{P(B)}}$ |

TAB. 7 – *Mesures déduites de la confiance par transformée affine*

La corrélation de points utilise le coefficient de corrélation de Pearson $r$ entre deux *itemsets* pour évaluer la force de leur liaison, qui peut être positive (se reporter à la table des exemples et contre-exemples de $A \Leftrightarrow B$) ou négative (se reporter à la table de $A \Leftrightarrow \overline{B}$). Le coefficient $r$ s'écrit :

$r(A,B) = \frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = \frac{\sqrt{P(A)}}{\sqrt{P(\overline{A})P(B)P(\overline{B})}} [Conf(A \rightarrow B) - P(B)]$

Celui-ci se simplifie en $r(A,B) = \frac{P(AB)-P(B)^2}{P(B)P(\overline{B})} = \frac{P(B)}{P(\overline{B})} [Conf(A \rightarrow B) - P(B)]$, lorsque $A$ et $B$ ont même distribution marginale ($P(A) = P(B)$), et en $r(A,B) = 2Conf(A \rightarrow B) - 1$, lorsque de plus cette distribution est équilibrée ($P(A) = P(B) = 0.5$). On constate ainsi que $r$ et $c$ sont équivalents lorsque l'on travaille sur un tableau dont les marges sont équilibrées, ce qui est à rattacher au fait que le tableau est alors symétrique, impliquant que toutes les règles covariantes ont la même confiance, de même que toutes les règles contravariantes.

## 6.2 Autres mesures

Parmi les mesures qui ne se réduisent pas à une transformée affine de la confiance, on citera notamment le $\chi^2$, la conviction, la *J-mesure* et les mesures construites à partir de l'indice d'implication, qui prennent toutes en compte les contre-exemples. On peut y ajouter la mesure de Sebag-Schoenauer, déjà évoquée, tout en rappelant que celle-ci est une transformation monotone croissante de la confiance.

Le $\chi^2$, qui est la statistique du test d'indépendance entre $A$ et $B$, utilisé notamment par Brin et al. (1997a), et le $\chi^2$ normalisé, noté $\varphi^2$, sont des fonctions du coefficient de corrélation de points $r(A,B)$ qui ne dépendent pas des probabilités de $A$ et de $B$. C'est ainsi que dans une même base, elles classent comme $r^2(A,B)$ :

$r^2(A,B) = \varphi^2 = \frac{\chi^2}{n} = \frac{[P(AB)-P(A)P(B)]^2}{P(A)P(B)((1-P(A))(1-P(B))}$

On remarquera ainsi que sous l'hypothèse d'indépendance, $nr^2(A,B) \approx N(0,1)^2$. Toutes ces variantes ne permettent pas de distinguer le cas $A \rightarrow B$ du cas $A \rightarrow \overline{B}$ et sont symétriques. On préférera $r(A, B)$ qui au moins est signé.

La conviction (Brin 1997a) est en quelque sorte un analogue du lift appliqué aux contre-exemples :

$$Conv(A \rightarrow B) = \frac{P(\overline{B})}{P(\overline{B}/A)} = \frac{P(A)P(\overline{B})}{P(A\overline{B})} = l(A \rightarrow \overline{B})^{-1} = \frac{1-P(B)}{1-conf(A \rightarrow B)}$$

On peut exprimer la mesure de Loevinger comme une fonction monotone croissante de la conviction :

$Loev(A \rightarrow B) = 1 - Conv(A \rightarrow B)^{-1}$

La mesure de Sebag et Schoenauer s'écrit en fonction de la confiance de $A \rightarrow B$ et de celle de $A \rightarrow \overline{B}$ :

$$Seb(A \rightarrow B) = \frac{conf(A \rightarrow B)}{conf(A \rightarrow \overline{B})} = \frac{conf(A \rightarrow B)}{1-conf(A \rightarrow B)} = \frac{1}{P(\overline{B})}\left(Conv(A \rightarrow B) - 1\right)$$

La *J-mesure* (Goodman et Smyth 1988) tient compte à la fois de la généralité de la règle et de sa capacité de prédiction. La généralité de la règle est mesurée par $P(A)$, alors que sa capacité de prédiction est évaluée par :

$$j(A \rightarrow B) = P(B/A)\log\left(\frac{P(B/A)}{P(B)}\right) + P(\overline{B}/A)\log\left(\frac{P(\overline{B}/A)}{P(\overline{B})}\right)$$

La *J-mesure* s'écrit ainsi :

$$\begin{aligned} J(A \rightarrow B) &= P(A) \times j(A \rightarrow B) \\ &= P(AB)\log\left(\frac{P(AB)}{P(A)P(B)}\right) + P(A\overline{B})\log\left(\frac{P(A\overline{B})}{P(A)P(\overline{B})}\right) \end{aligned}$$

### 6.3 Quelle mesure choisir ?

Notre but est de permettre un choix judicieux des mesures utilisées pour évaluer l'intérêt des règles d'association, compte tenu du but poursuivi et de la taille de la base de données. Dans cette perspective, une assistance au choix des mesures du type de celle proposée par Lenca, Meyer, Vaillant et Picouet dans le cadre du groupe GaFo-Qualité (Lenca et al. 2002) présente un grand intérêt.

Une fois choisi un panel de mesures appropriées, on peut conserver la partie extraction des fréquents des algorithmes du type Apriori (en $O(n)$ si on ne tient pas compte du nombre d'items) tout en substituant ce panel à la confiance lors de la sélection des règles, ce qui suppose que l'on se limite aux règles de fort support. On peut aussi renoncer au primat de la condition de support et rechercher directement les règles intéressantes selon le panel de mesures, ce qui oblige à se limiter à des règles simples.

## 7 Validation des règles : apports de la théorie de l'apprentissage statistique

En tout état de cause, que l'on utilise des mesures statistiques pour sélectionner les règles significatives, ou que l'on utilise des mesures descriptives pour classer les règles, on a besoin de valeurs critiques pour ces mesures, afin d'éliminer les règles sans intérêt. Dans un cadre probabiliste, pour déterminer les règles dont on est sûr à un risque donné qu'elles présentent pour chaque critère considéré une valeur supérieure

au seuil correspondant, on peut construire des intervalles de confiance asymptotiques unilatéraux à gauche pour les valeurs théoriques. A cet effet, on considère que les valeurs empiriques sont les estimateurs du maximum de vraisemblance des valeurs théoriques et on utilise la méthode delta (Goodman et Kruskal 1972) pour calculer les variances asymptotiques des estimateurs du maximum de vraisemblance et en déduire les intervalles de confiance. En procédant de la sorte, compte tenu du nombre de règles évaluées et du nombre de critères, on ne contrôle plus le risque réel auquel on opère. La théorie de l'apprentissage statistique apporte une solution originale à ce problème. Dans cette section, nous partons d'un précédent travail (Teytaud et Lallich 2001) pour proposer différents algorithmes utilisant les outils de l'apprentissage statistique (Vapnik 1995, Devroye et al. 1997) qui donnent des bornes uniformes pour toutes les règles et pour tous les critères considérés. Nous indiquons un premier algorithme qui nous servira de référence (algorithme $B$). Nous proposons ensuite diverses améliorations.

## 7.1 Algorithme de référence

Une approche simple des règles consiste à utiliser l'algorithme $B$ décrit ci-dessous :

1. Définir un ou plusieurs critères $c_1$,$c_2$,$c_3$,...,$c_N$. Pour plus de clarté, on se limitera par la suite à un seul critère $c$, mais l'extension à $N$ critères ne pose pas de problème.
2. Définir un ensemble $R$ de règles parmi lesquelles on va choisir des règles souhaitées pertinentes.
3. Se procurer un ensemble $T$ de transactions sur lesquelles on va tester les règles. On supposera par la suite que $T$ est indépendant et identiquement distribué (*iid*). On peut étendre facilement au cas aléatoire simple, d'autres extensions sont possibles.
4. On définit $c'(r)$, pour toute règle $r$, comme étant l'évaluation empirique du critère $c$ sur la règle $r$.
5. Utiliser un algorithme (quelconque) permettant d'extraire le sous-ensemble $E$ inclus dans $R$ des règles $r$ telles que $c'(r) > S$, avec $S$ un seuil donné.

Le défaut de cette approche est lié à la validation statistique. Rien ne garantit ici que $c(r) > S$, on sait seulement que $c'(r) > S$.

Une première solution serait de remplacer l'étape 5.) par l'étape 5'.), définissant l'algorithme $B'$:

5'. Successivement :

- pour chaque règle $r$, évaluer l'intervalle de confiance unilatéral $[a(r), +\infty]$, au seuil de confiance $1 - \delta$. C'est-à-dire que pour tout $r$, la probabilité pour que $c(r)$ ne soit pas supérieur à $a(r)$ est inférieure à $\delta$.
- fournir à l'utilisateur toutes les règles telles que $a(r) \geq S$.

S'il y a $k$ règles, la probabilité pour que dans un cas au moins, $c(r)$ sorte de l'intervalle n'est plus majorée que par $k\delta$. Finalement, la probabilité, pour une règle appartenant à $E$, d'avoir effectivement $c(r) > S$, n'est absolument pas contrôlée dès que $k$ devient trop grand.

Des solutions existent. La plus simple est la correction de Bonferroni. Si l'on a 50 règles, on divise le seuil de risque par 50 ; par suite toutes les règles $r$ dans $E$ seront effectivement garanties vérifier $c(r) > S$, avec une probabilité au moins $1 - \delta$. Le problème est que l'intervalle de confiance n'a alors plus de sens, $k$ étant généralement très grand. Il faut donc utiliser des résultats plus complexes.

## 7.2 Algorithme *VC*

On a notamment le résultat suivant, classique en théorie de l'apprentissage :

La probabilité pour que $|c(r) - c'(r)| > \varepsilon$ pour au moins une règle $r$ est majorée par :

$\delta = 8exp(mH(V/m))exp(-m\varepsilon^2/32)$, si $V > 2$ et $m > 2V$

où $V$ est la *VC-dimension* de l'ensemble $c(R)$ des $c(r)$, $n$ le cardinal de l'ensemble $T$ des transactions (on suppose ici les données *iid*), $H$ la fonction d'entropie définie par $H(x) = -xln(x) - (1-x)ln(1-x)$.

Ce résultat fournit une meilleure évaluation dans un grand nombre de cas (précisément, si $Ln\,|R|$ est suffisamment grand devant la *VC-dimension*, ce qui est notamment le cas si $R$ est infini et la *VC-dimension* finie). Des versions unilatérales de cette borne permettent de borner la probabilité pour que $c'(r) - c(r) > \varepsilon$, ce qui est notre problème.

Les résultats de type *VC-dimension* ont parfois mauvaise presse car on les accuse de conduire à des seuils de risque trop prudents. Cela est vrai, mais dans le cas des règles $n$ est souvent très grand, et ce résultat est asymptotiquement meilleur que la correction de Bonferroni. En outre, on peut ici s'attaquer à des ensembles infinis de règles sans difficulté; par exemple, dans le cas des transactions d'un supermarché, s'interroger sur le fait que le montant des achats de saucisses soit supérieur à $x$ euros, et non seulement sur le fait que des saucisses aient été achetées, ou d'autres discrétisations plus fines. Nous avons fait le calcul de la *VC-dimension* dans un grand nombre de cas (Teytaud et Lallich 2001). En fixant $\delta$ et $m$ et en résolvant l'équation, on obtient une précision $\varepsilon$. On peut alors appliquer l'algorithme qui suit, consistant à remplacer 5.) par 5".) :

5". Sélectionner toutes les règles $r$ dans $R$ telles que $c'(r) > S + \varepsilon$.

Avec l'algorithme *VC*, on peut affirmer, avec une confiance $1-\delta$, que toutes les règles sélectionnées vérifient bien le critère $c(r) > S$. On peut aussi exhiber en fait toutes les règles avec leurs intervalles de confiance (et par exemple montrer quelles sont les règles pour lesquelles on peut affirmer que $c(r) < S$). On gagnera parfois à utiliser d'autres bornes que celle proposée ici. Notamment on peut grandement tirer parti de critères entraînant de très petites ou très grandes fréquences. Par ailleurs, d'autres possibilités existent, parfois plus efficaces que la *VC-dimension* : les nombres de couverture et les coefficients de pulvérisation (Teytaud et Lallich 2001).

## 7.3 Exemple d'application par l'algorithme VC

Par la suite, nous noterons $C$ l'ensemble des fonctions caractéristiques (tableau 8) dont l'espérance doit être évaluée. C'est ainsi que pour un ensemble $R$ de règles dont

| Mesure d' intérêt | Fonction | Précision |
|---|---|---|
| Confiance | { AB, A } | $\epsilon \frac{\widehat{E(AB)}+\widehat{E(A)}}{\widehat{E(A)}^2-\epsilon\widehat{E(A)}}$ |
| Support | { AB } | $\epsilon$ |
| Lift | { AB,A,B} | $\epsilon \times \frac{\widehat{E(A)}\widehat{E(B)}+\widehat{E(AB)}\times(\widehat{E(A)}+\widehat{E(B)}+\epsilon)}{\widehat{E(A)}\widehat{E(B)}\times(\widehat{E(A)}\widehat{E(B)}+\epsilon\widehat{E(A)}+\epsilon\widehat{E(B)}+\epsilon^2}$ |
| *J-mesure*, $\phi^2$, $r$, $Z$ conviction | { AB,A,B} | voir méthode générale |

TAB. 8 – *Fonctions caractéristiques dont l'espérance doit être évaluée, et précision résultante sur l'intérêt de la règle pour la mesure d'intérêt considérée lorsque les espérances requises sont évaluées à $\epsilon$ près.*

on ne veut borner que la confiance, on a $C = \cup_{(A\Rightarrow B)\in R}\{AB,A\}$. Détaillons le principe du tableau 8 sur des exemples précis:

• La première colonne concerne la valeur que l'on cherche à mesurer. Par exemple, dans le cas du support, $C$ est l'ensemble des $AB$ trouvés dans les règles $A \rightarrow B$ (si l'on s'intéresse à plusieurs caractéristiques en même temps, il faut utiliser la réunion des $C$ proposés par chaque ligne). On applique alors les formules fournies par les équations comme l'équation fournie plus haut, qui nous donnent une précision $\epsilon$, bornant, avec le risque $\delta$ choisi, la différence entre l'évaluation empirique et la probabilité réelle d'occurrence, pour chacun des évènements de $C$. Dans le cas du support, la traduction est immédiate: on dispose directement de bornes sur $P(AB)$.

• Le cas de la confiance est plus complexe. On procède de même jusqu'à obtenir des intervalles de confiance sur $P(AB)$ et $P(A)$: $P(AB) \geq \widehat{E(AB)} - \epsilon$ et $P(A) \leq \widehat{E(A)} + \epsilon$. Alors $P(AB)/P(A) \geq \frac{\widehat{E(AB)}-\epsilon}{\widehat{E(A)}+\epsilon}$ (si le dénominateur est positif). Détailler la différence entre ce quotient et $P(AB)/P(A)$ amène la formule de la troisième colonne: si $E(AB)$ et $E(A)$ sont approximées par $\widehat{E(AB)}$ et $\widehat{E(A)}$ avec la précision $\epsilon$, alors $P(B|A)$ est approximée par $\frac{\widehat{E(AB)}}{\widehat{E(A)}}$ avec la précision $\epsilon\frac{\widehat{E(AB)}+\widehat{E(A)}}{\widehat{E(A)}^2-\epsilon\widehat{E(A)}}$. La troisième colonne du tableau donne l'erreur maximale commise en évaluant la confiance par $\widehat{E(AB)}/\widehat{E(A)}$. Des calculs similaires conduisent aux autres formules de la colonne.

• Si l'on s'intéresse à d'autres mesures, la formule devient plus compliquée, comme en témoigne la troisième ligne. Dans ces cas-là, le plus compréhensible est d'utiliser la "méthode générale" détaillée ci-dessous.

La méthode générale consiste à écrire que $E(D)$ est compris entre $\widehat{E(D)} - \epsilon$ et $\widehat{E(D)} + \epsilon$ dès lors que $D$ est dans $C$. D'autres bornes peuvent être envisagées lorsque l'espérance est petite (Teytaud et Lallich 2001). Ces bornes unilatérales permettent de borner supérieurement $E(D)$ par $\widehat{E(D)} + \epsilon$, avec $\epsilon$ plus petit (décroissant avec $1/n$ au lieu de $1/\sqrt{n}$).

## 7.4 Algorithme *HO* (comme *hold-out*)

D'autres solutions peuvent être proposées ; par exemple celle qui suit, fondée sur la technique du *hold-out* :

1. Couper la base $T$ en une base $T_1$ et une base $T_2$.
2. Définir $c"$ et $(c")'$ comme en 4. (algorithme B), mais pour les bases $T_1$ et $T_2$ respectivement.
3. Extraire dans $E_1$ les règles $r$ telles que $c"(r) > S$ (éventuellement, n'en garder qu'un sous-ensemble).
4. Définir des intervalles de confiance unilatéraux $[a(r), +\infty]$ pour $c(r)$, fondés sur $(c")'(r)$.
5. Garder seulement dans $E$ les règles $r$ de $E_1$ telles que $a(r) > S$.

Le risque est ici multiplié par le cardinal de $E_1$, à moins d'effectuer une correction de Bonferroni (en utilisant à l'étape 4 un risque divisé par le cardinal de $E_1$). Là aussi, pour peu que $Ln\,|E_1|$ soit suffisamment grand devant sa *VC-dimension*, on gagnera en précision en faisant intervenir la *VC-dimension*.

## 7.5 Algorithme *HOVC*:

Seule l'étape 4.) de l'algorithme *HO* est modifiée. On la remplace par l'étape 4'.) définie ci-dessous.

4'. Définir des intervalles de confiance $[a'(r), +\infty]$ pour $c(r)$, grâce à la formule de *VC-dimension* plus haut, avec V la *VC-dimension* de l'ensemble des $c(r)$ pour $r$ dans $E_1$.

On peut s'inquiéter du risque de se retrouver avec $E$ très petit. Néanmoins, il faut bien voir, sur la formule plus haut, que le risque est en exponentielle décroissante du nombre de transactions divisé par la *VC-dimension*; ou que la précision $\varepsilon$ décroît comme l'inverse de la racine du nombre de transactions. Pour un grand nombre de transactions, et cela est souvent le cas, on peut raisonnablement utiliser ces techniques. Une difficulté de l'algorithme *HOVC* est liée au calcul de la *VC-dimension*.

Pour un ensemble de règles extrait de manière aléatoire comme dans l'algorithme *HOVC*, le calcul de *VC-dimension* peut s'avérer beaucoup plus difficile. Une borne peut-être utilisée (basée sur la finitude de $E_1$ par exemple), mais on risque alors de devenir peu efficace. Lorsque le nombre de transactions est trop petit pour ces techniques, on peut envisager le recours au *bootstrap*.

## 7.6 Algorithme *BS*

On propose l'algorithme ci-dessous :

1. Définir $c'(r)$, l'évaluation de $c(r)$ sur l'ensemble $T$ de transactions.

2. Un grand nombre de fois :

   - Tirer au sort, avec remise, une liste $T'$ d'éléments de $T$, de même cardinal de $T$. Cela implique donc que, en général, des éléments de $T$ puissent apparaître plusieurs fois dans la liste $T'$.

   - Définir $c"(r)$, l'évaluation de $c(r)$ sur la liste $T'$ de transactions.

   - Calculer $\varepsilon$, le supremum des écarts $c"(r) - c'(r)$ (et non $|c"(r) - c'(r)|$) pour $r$ dans $R$.

3. On obtient donc un grand nombre de valeurs $\varepsilon$. Examiner $\varepsilon(\delta)$, le quantile $(1-\delta)$ de ces $\varepsilon$.

4. Garder dans $E$ toutes les règles $r$ de $R$ telles que $c'(r) > S + \varepsilon(\delta)$.

La justification de cet algorithme est la suivante. On suppose que l'ensemble des $c(r)$ est *Donsker*. Ceci est vrai dès que le cardinal est fini ou dès que la *VC-dimension* est finie ; en fait, c'est vrai dans un cadre beaucoup plus général et ne pose pas de problème pour toutes les familles usuelles de règles. Si la distribution de l'ensemble des transactions est absolument continue par rapport à la mesure de *Lebesgue*, on peut même travailler sur des classes *non-Donsker*, pourvu que, essentiellement, la famille soit *prégaussienne* (Radulovic et Wegkamp 2000).

Alors, le processus des $\sqrt{n}(c"(r)-c'(r))$ est un estimateur asymptotiquement consistant du processus des $\sqrt{n}(c'(r) - c(r))$.

Les théorèmes liés au *bootstrap* garantissent donc que $\varepsilon(\delta)$ converge vers le supremum sur $r$ de $c'(r) - c(r)$ faiblement en $o(1/\sqrt{n})$, avec $n$ le nombre de transactions. L'écart avec la théorie de la *VC-dimension* (ou les autres théories non-asymptotiques) est donc dans ce $o(.)$ au lieu d'un $O(.)$. Finalement, le risque de première espèce (le risque d'accepter une règle qui en fait ne vérifie pas le critère) est moins bien garanti (le résultat est asymptotique), mais le risque de seconde espèce (le risque d'oublier d'accepter une règle valide) est largement meilleur. On évite ainsi le défaut traditionnel reproché à la théorie VC.

On peut affirmer ainsi, avec confiance environ $1 - \delta$, que toutes les règles extraites par l'algorithme *BS* vérifient bien $c(r) > S$. Le point qui peut fâcher est le « environ » (que l'on n'a pas en *VC-théorie*). Il s'agit en effet d'une estimation asymptotique. On peut simplement ajouter qu'en pratique le *bootstrap* est en général très efficace.

## 7.7 Algorithme HOBS

On peut imaginer *HOBS*, équivalent de *HOVC* utilisant le *bootstrap* au lieu des bornes *VC*, lors de l'étape sur $T_2$. Un tel algorithme présente le même défaut que *BS* (présence du « environ ») au niveau de la rigueur du résultat, mais gagne grandement en vitesse par rapport à VC, en simplicité mathématique (pas de calcul de *VC*-dimension) par rapport à *VC* ou *HOVC*, et en précision par rapport à *VC* ou *HOVC* (en raison du $o(.)$ au lieu du $O(.)$, comme *BS*). Il pose quelques problèmes (choix des tailles de $T_1$ et $T_2$, complexité de programmation).

### 7.8 Préconisations

Nous avons présenté un certain nombre d'algorithmes issus de la théorie de l'apprentissage statistique qui permettent de proposer des bornes uniformes pour tous les critères de qualité envisagés et toutes les règles extraites. En ce qui concerne le *bootstrap*, on notera que le nombre de tirages doit être suffisant pour assurer une précision de l'ordre de la finesse du seuil de risque choisi. Les seuils usuels étant petits, ce nombre est rapidement de l'ordre de quelques dizaines ou centaines de milliers, sauf à ajouter à l'imprécision due à l'approximation asymptotique en le nombre d'individus un terme d'imprécision sur le risque dû à l'approximation asymptotique en le nombre de rééchantillonnages. Nous proposons donc :

- l'algorithme $VC$ si les bases sont grandes ($VC$ demande de grandes bases) ;
- l'algorithme $BS$ si les bases sont petites ($BS$ est coûteux en temps de calcul) ;
- l'algorithme $HOVC$ si les bases sont de taille intermédiaire;
- lorsque l'ensemble $c(R)$ a une *VC-dimension* trop grande pour que $VC$ ait un risque de seconde espèce acceptable (ie si $VC$ ne sélectionne pas un nombre de règles acceptable), si l'ensemble $E_1$ est trop grand pour que $HOVC$ ait un risque de seconde espèce acceptable (ie si $HOVC$ lui-même ne sélectionne pas un nombre de règles suffisant), si la base est trop grande pour que $BS$ soit utilisé, alors $HOBS$ s'impose.

## 8 Conclusion et travaux futurs

Dans ce papier, qui doit beaucoup à la réflexion engagée dans le cadre de l'action GaFo-Qualité, après avoir évoqué l'intérêt et les limites des seuls critères de support et de confiance pour apprécier l'intérêt des règles d'association, nous avons voulu donner à l'utilisateur les moyens de choisir d'autres critères qui prennent en compte la nature particulière des règles d'association tout en étant les plus adaptés au problème traité.

Une multiplicité de tests doit alors être pratiquée : pour chaque règle, le test d'indépendance de l'antécédent et du conséquent puis le test du seuil retenu pour chaque critère. Face à la difficulté de contrôler le risque de cette multitude de tests, différents algorithmes issus de la théorie de l'apprentissage statistique ont été proposés qui construisent des bornes uniformes pour tous les critères de qualité envisagés et toutes les règles extraites.

Diverses améliorations sont envisageables, notamment le choix assisté des mesures (Lenca et al. 2002), ainsi que la prise en compte de la dimension « extraction » et la gestion des règles obtenues. Il nous semble que le classement des règles par différents critères et la recherche des règles qui dominent les autres fournit une piste intéressante.

## Références

Agrawal R. et Srikant R. (1994), Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994.

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

Azé J. et Kodratoff Y. (2002), Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In n° spécial revue ECA, *Actes Colloque EGC 2002*, Montpellier, pp. 143-154.

Bastide Y., Taouil R. Pasquier N., Stumme G. et Lakhal L. (2002), PASCAL : un algorithme d'extraction des motifs fréquents. *Technique et Science Informatique*, vol. 21, 1, pp. 65-95.

Bayardo R.J. (1998), Efficiently mining long patterens from databases, *Proc. ACM SIGMOD'98*, pp. 85-93, Seattle, Washington, USA.

Becquet C., Blachon S., Jeudy B., Boulicaut J.-F. et Gandrillon O. (2002), Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data, *Genome Biology*, 2002, 3(12).

Boulicaut J.-F. et Bykowski A. (2000a), Frequent closures as a concise representation for binary data mining, *Proc. of the 4th PAKDD Conf.*, april 2000, pp. 62-73.

Boulicaut J.-F., Bykowski A. et Rigotti C. (2000b), Approximation of frequency queries by mean of free-sets, *Proc. of the 4th PKDD Conf.*, sept. 2000, pp. 75-85.

Brin S., Motwani R. et Silverstein C. (1997a), Beyond market baskets : generalized associations rules to correlations. In *Proceedings of ACM SIGMOD'97*, 1997.

Brin S., Motwani R., Ullman J.-D. et Tsur S. (1997b), Dynamic itemset counting and implication rules for market basket data. *SIGMOD Record* (ACM Special Interest Group on Management of Data), 26(2):255, 1997.

Czekanowski J. (1913), *Zarys metod statystycznych (Die Grundzuge der statischen Metoden)*, Warsaw.

Devroye L., Györfi L. et Lugosi G. (1997), *A probabilistic theory of pattern recognition*, Springer, 1997

Freitas A. (2000), Understanding the crucial difference between classification and discovery of association rules - a position paper. *SIGKDD Explorations*, vol. 2, 1, pp. 65-69, 2000.

Goodman L.A et Kruskal W.H. (1972), Measures of association for cross-classification, IV, Simplification of asymptotic variances, *JASA*, 1972, 67, pp. 415-421.

Goodman R.M. et Smyth P. (1988), Information theoretic rule-induction, *Proc. of ECAI'88*, 1988.

Gras R. (1979), *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathematiques*, Thèse d'Etat, Université Rennes 1, 1979.

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

Gras R. et Lahrer A. (1993), L'implication statistique : une nouvelle méthode d'analyse des données, *Mathématiques Informatique et Sciences Humaines*, 120, pp. 5-31, 1993.

Guillaume S. (2000), *Traitement des données volumineuses, mesures et algorithmes*

*d'extraction de règles d'association et règles ordinales*, Thèse de doctorat, Université de Nantes, 2000.

Hajek P. et Rauch J. (1999), Logics and statistics for association rules and beyond, Tutorial PKDD'99, Prague, 1999.

Hajek P., Havel et Chytil (1966), The GUHA method of automatic hypotheses determination, *Computing*, 1, pp. 293-308, 1966.

Han J., Pei J. et Yin Y. (2000), Mining frequent pattern without generation candidate. In *Proceedings of the 2000 ACM SIGMOD*, Dallas, Texas, USA, 2000.

Hipp J., Guntzer U. et Gholamreza N. (2000), Algorithms for association rule mining - a general survey and comparison. *SIGKDD Explorations*, vol. 2, 1, pp. 58-64, 2000.

Kodratoff Y. (1999), Quelques contraintes symboliques sur le numérique en ECD et en ECT, Ecole Modulad/SFdS-Inria, 1999.

Kulczynski S. (1928), Die Pflanzenassoziationen der Pieninen, *Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat.*, Ser. B, Suppl. II (1927), pp. 57-203.

Lallich S. (2002), Mesure et validation en extraction des connaissances à partir des données, Habilitation à diriger les recherches, Université Lyon 2, 2002.

Lenca P., Meyer P., Vaillant B. et Picouet P. (2002), Aide multicritère à la décision pour évaluer les indices de qualité des connaissances, *Conférence EGC 03, Extraction et Gestion des Connaissances, Revue RSTI, série RIA-ECA*, vol. 17, n° 1-2-3, pp. 271-282, 2003.

Lerman I.C. (1988), Comparing partitions, mathematical and statistical aspects, *Classification and related methods of data analysis*, H. H. Bock (ed), Elsevier Science Publishers, 1988.

Lerman I.C. et Azé J. (2002), Indice Probabiliste Discriminant (de vraisemblance du lien) d'une règle d'association en cas de très grosses données, *Contribution au rapport d'activité du Groupe Qualité de l'action GaFo-Données*, 2002.

Lerman I.C., Gras R. et Rostam H. (1981), Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines*, n° 74, pp. 5-35 et n° 75, pp. 5-47, 1981.

Loevinger J. (1947), A systematic approach to the construction and evaluation of tests of ability, *Psychological monographs*, 61, n° 4, 1947.

Ochiai A. (1957), Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bull. Jpn. Soc. Sci. Fish.*, 22, pp. 526-530.

Pasquier N., Bastide Y., Taouil R. et Lakhal L. (1999a), Discovering frequent closed itemsets for association rules, *Proc. of the 7th ICDT, Int'l Conference on Database Theory*, n° 1540 LNCS, Springer, jan. 1999, pp. 398-416.

Pasquier N., Bastide Y., Taouil R. et Lakhal L. (1999b), Efficient Mining of association rules using closed itemsets lattices, *Journal of Information Systems*, vol. 24, 1, pp. 25-46, Elsevier Science.

Pearl J. (1988), *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann, 1988.

Piatetsky-Shapiro G. (1991), Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pp. 229-248. AAAI Press / The MIT Press, 1991.

Radulovic D. et Wegkamp M. (2000), Weak convergence of smoothed empirical processes : beyond Donsker classes. *High Dimensional probability II*, E. Gine, D. Mason, J. Wellner, Eds, Birkhauser, 2000.

Russell P. F. et Rao T. R. (1940), On habitat and association of species of anopheline larvae in south-eastern Madras, *J. Malar. Inst. India*, 3, pp. 153-178.

Savasere A., Omiecinski E. et Navathe S. (1995), An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st Conference on Very Large Databases - VLDB'95*, pp. 432-444, Zurich, Switzerland, 1995.

Sebag M. et Schoenauer M. (1988), Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases, in J. Boose, B. Gaines, M. Linster Eds, *Proc. of the European Knowledge Acquisistion Workshop*, EKAW'88, pp. 28-1-28-20, Gesellschaftfür mathematik und Datenverarbeitung mbH, 1988.

Srikant R. et Agrawal R. (1995), Mining generalized associations rules. In *Proc. of the 21st Int. Conf. on Very Large Databases*, VLDB'95, Zurich, Switzerland, 1995.

Tan P.N., Kumar V., Srivastava J. (2002), Selecting the right interestingness measure for association patterns. In *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 32-41.

Teytaud O. et Lallich S. (2001), Bornes uniformes en extraction de règles d'association, *Actes Colloque CAp 2001*, Grenoble, pp. 133-148, 2001. Version étendue disponible à url = "citeseer.nj.nec.com/477625.html".

Toivonen H. (1996), Sampling large databases for association rules, *Proc. 22nd VLDB Conference*, pp. 134-145,Bombay, Indes.

Vapnik V.N. (1995), *The nature of statistical learning*, Springer, 1995.

Zaki M.J., Parthasarathy S., Ogihara M. et Li W. (1997), New algorithms for fast discovery of associations rules. In *Proceedings of 1997 ACM SIGMOD* Int'l Conference on KDD and Data Mining, KDD'97, Newport Beach, Californie, 1997.

Zhang T. (2000), Association rules. In T. Terano, H. Liu, A.L.P. Chen (Eds), *Actes Conférence PAKDD 2000*, LNAI 1805, pp 245-256, Springer-Verlag, 2000.

## Summary

The research of interesting associations rules in databases is an important task of knowledge data discovery. Algorithms based on support and confidence, such as Apriori, brought a neat solution to the rules extraction problem. As shown in this article, these algorithms miss interesting rules and some of the rules they select are of no interest. Furthermore, they produce too many rules. It is therefore important to have other measures to complement support and confidence. In this article, we review the different measures suggested in the literature and we propose criteria for their evaluation. We then suggest a validation method using tools issued from the statistical learning theory, notably *VC-dimension*. Facing the high number of measures and the multitude of candidate rules, these tools enable to set uniform non asymptotic bounds for all rules and measures simultaneously.