Évaluation de la Résistance au Bruit de quelques Mesures Quantitatives

Jérôme Azé*, Sylvie Guillaume**, Philippe Castagliola***

*CNRS, LRI - Université Paris Sud - 91405 Orsay Cedex
Jerome.Aze@lri.fr
http://www.lri.fr/~aze
**Laboratoire LIMOS, UMR 6158 CNRS
Université Blaise Pascal - Complexe scientifique des Cézeaux
63177 Aubiere Cedex
Sylvie.Guillaume@isima.fr
***École des Mines de Nantes / IRCCyN
4, rue Alfred Kastler - 44307 Nantes Cedex 3
Philippe.Castagliola@emn.fr

Résumé. L'extraction de connaissances dans des données réelles est difficile car les données sont rarement parfaites. L'étude de l'impact du bruit contenu dans les données sur la qualité des résultats obtenus permet de mieux comprendre le comportement des mesures de qualité. Dans cet article, nous présentons différentes mesures quantitatives permettant d'extraire des connaissances dans les données. Pour chacune de ces mesures, une étude empirique de l'impact d'un bruit relativement réaliste sur une base de données bancaires est réalisée. Le comportement des différentes mesures en présence des données bruitées permet d'établir un critère de qualité supplémentaire. Ce nouveau critère lié à la sensibilité des mesures aux données bruitées permet de mieux contrôler le choix des mesures lors du processus d'extraction des connaissances.

1 Introduction

De nombreuses études (Agrawal et al. 1996 ; Mannila et al. 1994 ; Srikant et Agrawal 1996 ; Park et al. 1995 ; Pasquier 2000) ont été réalisées sur la recherche d'algorithmes efficaces d'extraction de règles d'association pour des données discrètes. Pour les autres types de données, un codage disjonctif complet, précédé pour les variables quantitatives d'une discrétisation, est nécessaire afin d'utiliser ces algorithmes (Srikant et Agrawal 1996). Cette transformation des données a deux inconvénients, tout d'abord une augmentation du nombre de règles extraites et ensuite une perte de la structure d'ordre des variables ordinales (variables qualitatives ordinales et variables quantitatives). Afin de remédier à ce problème et de pouvoir extraire des associations directement sur les variables quantitatives sans avoir à les transformer, une étude de différentes mesures quantitatives (coefficient de corrélation linéaire, mesure de vraisemblance du lien, intensité d'implication de A. Larher, intensité de propension et intensité d'inclination) a été effectuée (Guillaume et Castagliola 2003). Cette étude s'est intéressée à deux problématiques, tout d'abord au comportement de ces mesures dans plusieurs situa-

tions à savoir l'indépendance entre les variables, l'incompatibilité, l'implication logique, et aux situations intermédiaires et pour finir au comportement de ces mesures dans le cas de données volumineuses, caractéristique importante de l'extraction de connaissances. Comme ces mesures sont destinées à travailler avec des données réelles qui sont rarement parfaites et peuvent couramment présenter du bruit, il est important de vérifier qu'elles restent efficaces en présence de données bruitées. Nous devons donc connaître l'influence du bruit sur les associations extraites et plus particulièrement sur le pourcentage d'associations pertinentes qui disparaissent ainsi que sur le pourcentage d'associations non pertinentes qui apparaissent. Une telle étude a déjà été réalisée (Azé et Kodratoff 2002) sur des mesures binaires (support, confiance, dépendance et moindre-contradiction).

Ainsi, cet article est organisé de la façon suivante. Dans la section 2 nous présentons quelques mesures quantitatives permettant d'extraire des associations entre **des couples de** variables numériques et dans la section 3 nous étudions plus particulièrement une forme de bruit altérant les données. Nous comparons dans la section 4 le comportement de ces différentes mesures en présence de données bruitées et nous terminons par une conclusion résumant l'ensemble des points abordés et par quelques perspectives envisagées pour la suite de ce travail.

2 Mesures Quantitatives

Dans cette section, nous présentons deux mesures quantitatives de similarité (le coefficient de corrélation linéaire significatif et la mesure de vraisemblance du lien) permettant de détecter des associations entre deux variables (X, Y) et une mesure implicative (l'intensité d'inclination) permettant de détecter des règles $(X_1...X_p \rightarrow Y_1...Y_q)$, c-à-d. des associations orientées entre p et q variables.. Dans la suite de l'article, nous utiliserons le terme association pour désigner des associations ou des règles d'association. Les deux mesures implicatives suivantes entre deux variables quantitatives, l'intensité de propension (Lagrange 1997) et l'intensité d'implication (Larher 1991) ne sont pas étudiées car l'intensité d'inclination en est une généralisation.

2.1 Coefficient de Corrélation Linéaire Significatif

Le coefficient de corrélation linéaire r est une mesure de liaison linéaire entre deux variables quantitatives X et Y. Lorsque r est proche de 0, les deux variables sont indépendantes ; lorsque r est proche de 1, les deux variables évoluent dans le même sens selon approximativement une droite de pente positive et pour finir lorsque r est proche de -1, les deux variables évoluent cette fois-ci en sens inverse selon approximativement une droite de pente négative. La caractéristique de l'extraction de connaissances à partir des données étant de travailler avec des données volumineuses, nous retenons l'approximation faite par Saporta dans (Saporta 1990), c'est-à-dire pour une population Ω de taille N supérieure à 100, la variable aléatoire R, dont le coefficient de corrélation r est une valeur observée, suit approximativement la loi normale de moyenne 0 et d'écart-type $\frac{1}{\sqrt{N-1}}$. Comme nous recherchons les liaisons significatives entre variables (*liaisons selon une droite de pente positive et liaisons selon une droite de pente négative*), nous

retenons l'indice significatif de dépendance suivant

$$CCLS = Pr(|R| \leq |r|)$$

2.2 Mesure de Vraisemblance du Lien

La mesure de vraisemblance du lien (Lerman 1981) évalue si le nombre des associations positives (*c'est-à-dire le nombre des individus vérifiant de fortes valeurs pour* X *et de fortes valeurs pour* Y) est significativement élevé comparativement à ce que l'on obtiendrait si X et Y étaient indépendantes. L'indice brut de similarité est défini par $v_0 = \sum_{i=1}^{N} x_i y_i$ où x_i et y_i sont respectivement les valeurs prises par les variables X et Y pour l'individu t_i. I.C. Lerman a démontré que la variable aléatoire V dont cet indice brut v_0 est une réalisation suit asymptotiquement la loi normale $\mathcal{N}(\mu, \sigma)$ de moyenne $\mu = N\mu_X\mu_Y$ et de variance $\sigma^2 = \frac{N^2}{N-1} v_X v_Y$ avec N le nombre de transactions[1], μ_X et μ_Y respectivement les moyennes des variables X et Y, et v_X et v_Y respectivement les variances des variables X et Y. De plus, il a démontré que l'indice brut normalisé $v_{on} = \frac{v_0 - \mu}{\sigma}$ est égal à $\sqrt{N-1}r$, où r est le coefficient de corrélation linéaire défini précédemment (voir section 2.1). La mesure de similarité locale $S_L(X,Y)$ est donc définie de la façon suivante :

$$S_L(X,Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{N-1}*r} e^{-\frac{1}{2}t^2} dt$$

Lorsque la valeur de r est négative (*respectivement positive*) et la taille de la population N importante, la valeur de $z = \sqrt{N-1}r$ tends vers moins l'infini (*respectivement plus l'infini*), et par conséquent la valeur de $S_L(X,Y)$ tends vers 0 (*respectivement 1*). Pour finir, lorsque r vaut 0, la valeur de $S_L(X,Y)$ est égale à $0,5$. Ainsi, dans le cas de données volumineuses, cette mesure locale n'est pas sélective car elle ne peut prendre que trois valeurs : 0, 1 et $0,5$. Afin de remédier à ce problème, I.C. Lerman a proposé une mesure de similarité globale $S_G(X,Y)$ définie de la façon suivante :

$$S_G(X,Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{z-\mu_z}{\sigma_z}} e^{-\frac{1}{2}t^2} dt$$

où μ_z et σ_z sont respectivement la moyenne arithmétique des valeurs de z extraites sur la population Ω et la variance de ces mêmes valeurs.

2.3 Intensité d'inclination

L'intensité d'inclination (Guillaume 2002) évalue si le nombre des individus ne vérifiant pas fortement la règle $X \rightarrow Y$ (*c'est-à-dire le nombre des individus vérifiant de fortes valeurs pour* X *et de faibles valeurs pour* Y) est significativement faible comparativement à ce que l'on obtiendrait si X et Y étaient indépendantes. Soient X et Y deux variables quantitatives prenant respectivement leurs valeurs x_i et y_i $(i = 1, .., N)$

1. ou observations ou encore individus.

dans les intervalles $[x_{min}..x_{max}]$ et $[y_{min}..y_{max}]$ et soit q_0 la mesure brute de non-inclination définie de la façon suivante $q_0 = \sum_{i=1}^{N}(x_i - x_{min})(y_{max} - y_i)$. Soient μ_X, μ_Y les moyennes arithmétiques de respectivement X et Y et v_X, v_Y les variances de X et Y. L'intensité d'inclination est donnée par la formule suivante :

$$\varphi(X \rightarrow Y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{q_0}^{+\infty} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt$$

avec $\mu = n(\mu_X - x_{min})(y_{max} - \mu_Y)$
$\sigma^2 = n[\sigma_X\sigma_Y + \sigma_Y(\mu_X - x_{min})^2 + \sigma_X(y_{max} - \mu_Y)^2]$

Cette mesure garantit que les effectifs observés s'écartent significativement des effectifs théoriques et particulièrement en présence de fortes valeurs pour X et faibles valeurs pour Y.

3 Données bruitées

Afin d'évaluer la résistance au bruit des mesures présentées dans la section 2, nous utilisons le protocole de test suivant :
Pour chacune des mesures, faire :

1. Calculer l'ensemble $\mathcal{A}$ des associations qui présentent une valeur supérieure ou égale à un seuil d'élagage donné.

2. Injecter du bruit dans la base de données,

3. Extraire le nouvel ensemble $\mathcal{A}'$ des associations à partir de la base de données bruitée.

4. Comparer les deux ensembles $\mathcal{A}$ et $\mathcal{A}'$ afin de calculer le nombre des associations qui ont disparu et le nombre des associations qui sont apparues.

Ce processus est répété k fois et à l'issue de ces répétitions, nous calculons la moyenne et l'écart-type du nombre des associations qui ont disparu et du nombre des associations qui sont apparues.

(Azé et Kodratoff 2002) ont étudié trois formes de bruit : (1) une seule variable est bruitée, (2) plusieurs variables sont bruitées et (3) une répartition aléatoire du bruit dans la base.
Ils ont montré que les deux dernières formes de bruit engendrent les plus mauvais résultats (Azé et Kodratoff 2002), c'est pourquoi nous avons décidé de tester les mesures quantitatives avec une de ces deux formes, nous avons retenu la dernière forme, c'est-à-dire une répartition aléatoire du bruit dans la base.

Avant d'expliquer la technique utilisée pour injecter du bruit dans plusieurs variables, nous définissons celle qui a été retenue pour injecter du bruit dans une seule variable.

Soit une base de données composée de N individus t_i $(i = 1, .., N)$ décrits par p variables quantitatives X_1, ... , $X = X_j$, .. , X_p $(j = 1, .., p)$ et soit x_i la valeur de la

variable X prise par l'individu t_i. Nous supposons que la variable X prend ses valeurs dans l'intervalle $[x_{min}..x_{max}]$.

Soient $F_X(x, a, b, c, d)$ la fonction de répartition de la variable X et $F_X^{-1}(y, a, b, c, d)$ la fonction inverse de F_X avec $x \in [x_{min}..x_{max}]$ et $y \in [0..1]$. Les paramètres a et b déterminent la forme de ces deux fonctions, le paramètre c correspond à la valeur minimale x_{min} de la variable X et le paramètre d représente le taux de couverture[2] pour cette valeur minimale ($X = x_{min}$).

Afin d'obtenir la valeur bruitée x'_i pour la valeur observée x_i, nous modifions les paramètres a, b et d en choisissant une nouvelle valeur a' (*respectivement* b' *et* d') dans l'intervalle $[a(1-\alpha), .., a(1+\alpha)]$ (*respectivement dans les intervalles* $[b(1-\alpha), .., b(1+\alpha)]$ *et* $[d(1-\alpha), .., d(1+\alpha)]$). Afin d'effectuer des tests réalistes, nous ne pouvons changer la valeur du paramètre c car sinon nous obtiendrions une nouvelle fonction de répartition trop différente de la fonction initiale. N'oublions pas que l'injection de bruit a pour but de simuler des erreurs dans la base et que cette nouvelle base doit être assez proche de la base d'origine. Ce choix est lié à la contrainte suivante : nous ne voulons pas modifier le domaine de variation de la variable bruitée mais juste la répartition de quelques valeurs (choisies aléatoirement dans l'intervalle $[x_{min}..x_{max}]$). En effet, nous pensons qu'il est raisonnable de considérer que le bruit, ayant pu altérer les données, n'a pas modifié le domaine de variation de celles-ci.

La nouvelle valeur x'_i pour l'individu t_i est égale à $F_X^{-1}(y_i, a', b', c, d')$ avec $y_i = F_X(x_i, a, b, c, d)$ comme le montre la Figure 1.

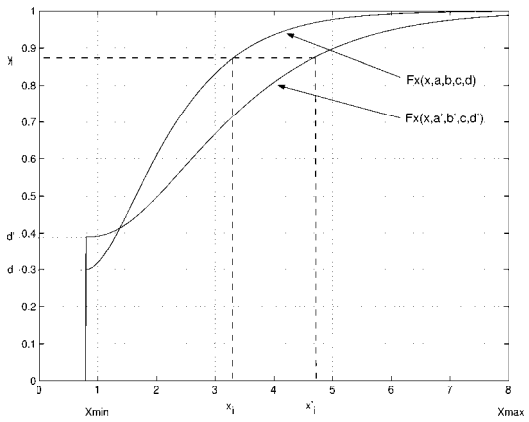

FIG. 1 – *Injection du bruit dans la base pour une variable X.*

Dans la Figure 1, $F_X(x, a, b, c, d)$ est la fonction de répartition gamma $F_X(x; 2; 0.8; 0.8; 0.3)$ et la valeur de α retenue est égale à 0.3.

Nous souhaitons introduire du bruit pour s ($s < p$) variables X_1 ,.. , $X = X_k$, .., X_s ($k = 1, .., s$). Soit γ le pourcentage de bruit injecté dans la base de données. Plus généralement, soit $x_i^{(k)}$ la valeur de la variable X_k pour l'individu t_i, nous allons

2. ou support.

modifier de façon aléatoire $\gamma * N$ valeurs $x_i^{(k)}$ par le processus précédent.

4 Evaluation sur des Données Bancaires

Dans cette section, nous présentons les tests effectués sur une base de données bancaires. Tout d'abord nous décrivons la base de données et ensuite dans la section 4.2 nous expliquons l'injection de bruit dans cette base. Pour finir, nous discutons et comparons l'effet du bruit sur les différentes mesures présentées en section 2.

4.1 Description de la Base

La base de données se compose de 47112 individus décrits par 44 variables quantitatives. Ces variables peuvent être classées en trois catégories:

1. les informations relatives aux clients (*âge, ancienneté*),
2. les différents produits financiers proposés par la banque (*actions, obligations, ...*),
3. les statistiques sur les différents comptes ouverts par les clients (*montant des ressources, montant des encours prêt, ...*).

Les variables relatives aux produits financiers proposés par la banque peuvent également être répertoriées en deux catégories:

(a) les variables mémorisant les encours de chaque produit financier pour tous les clients de la base,

(b) les variables comptabilisant le nombre de comptes ouverts par le client pour chaque produit financier.

4.2 Injection du Bruit dans la Base

Nous avons choisi d'introduire du bruit dans 41 variables. Nous avons éliminé la variable "ancienneté" car 19% des clients ont une ancienneté de 26 ans, environ trois fois plus que les autres valeurs (*la valeur minimale de cette variable est égale à* 1 *et la valeur maximale est égale à* 26). Lorsque la valeur de l'ancienneté pour un individu était inconnue, la valeur maximale était mémorisée. Les deux autres variables non retenues (*solde moyen des comptes à vue et total des ressources*) sont des variables de la troisième catégorie (*statistiques sur les comptes*) c'est-à-dire des combinaisons de variables de la deuxième catégorie. En effet, la fonction de distribution $F_X(x, a, b, c, d)$ de ces deux variables n'a pas été trouvée et une étude complémentaire doit être menée.

4.3 Définition et estimation de lois hybrides

Cette section décrit la technique qui a été utilisée pour trouver les fonctions de répartition des variables de la base de données bancaires.

Les échantillons correspondants aux 41 variables retenues pour l'étude peuvent se mettre sous la forme $\{x_{(i)}, f_{(i)}\}$, $i = 1, \ldots, n$, où $x_{(i)}$ sont des occurences ordonnées et

$f_{(i)}$ sont des fréquences. La principale caractéristique de ces données est d'avoir une fréquence $f_{(1)}$ très élevée ($f_{(1)} > 0.9$). Afin d'obtenir une modélisation paramétrique qui prenne en compte cette particularité, nous avons développé un ensemble de lois hybrides à quatre paramètres (a, b, c, d), avec $d \in [0, 1[$, basées sur des lois de probabilité classiques (gamma, lognormale et Weibull) à trois paramètres (a, b, c). Nous allons expliciter dans ce qui suit comment ces lois hybrides sont obtenues, comment générer des nombres aléatoires selon ces lois et comment estimer les paramètres (a, b, c, d). Soit X une v.a. continue définie sur $[c, +\infty[$, dont la fonction de répartition est $F_X(x, a, b, c)$ et vérifie $F_X(c, a, b, c) = 0$. On souhaite définir, à partir de la fonction de répartition $F_X(x, a, b, c)$ de la v.a. X, une nouvelle fonction de répartition $F_Y(y, a, b, c, d)$ définie sur $[c, +\infty[$, ayant pour propriété $F_Y(c, a, b, c, d) = d$. Pour cela, on propose de définir la fonction de répartition $F_Y(y, a, b, c, d)$ de la manière suivante

$$F_Y(y, a, b, c, d) = \{d + (1 - d)F_X(y, a, b, c)\}1_{y>c}$$

On voit clairement que $F_Y(y, a, b, c, d) = 0$ pour $y < c$, $F_Y(c, a, b, c, d) = d$ et $F_Y(y, a, b, c, d) = d + (1 - d)F_X(y, a, b, c)$ pour $y > c$. Puisque $F_X(c, a, b, c) = 0$, la distribution de probabilité de la v.a. Y est

$$f_Y(y, a, b, c, d) = (1 - d)f_X(y, a, b, c)1_{y\geq c} + d1_{y=c}$$

On montre facilement que le moment non centré d'ordre r de la v.a. Y est égal à $m_r(Y) = (1-d)m_r(Y)+dc^r$ et en "inversant" la fonction de répartition $F_Y(y, a, b, c, d)$, on obtient la fonction de répartition inverse définie pour $d < \alpha < 1$

$$F_Y^{-1}(\alpha, a, b, c, d) = F_X^{-1}\left(\frac{\alpha - d}{1 - d}, a, b, c\right)$$

On déduit donc que pour générer aléatoirement une v.a. Y de fonction de répartition $F_Y(y, a, b, c, d)$, il suffit de tirer une v.a. U uniforme sur $(0, 1)$ et de calculer

$$Y = \begin{cases} c & \text{si } U \leq d \\ F_X^{-1}\left(\dfrac{U - d}{1 - d}, a, b, c\right) & \text{si } U > d \end{cases}$$

A partir d'un échantillon $\{x_{(i)}, f_{(i)}\}$, $i = 1, \ldots, n$, on peut proposer comme estimateurs initiaux pour c et d, $\hat{c} = x_{(1)}$ et $\hat{d} = f_{(1)}$. Pour ce qui est de l'estimation des paramètres a et b nous allons étudier les trois cas suivants, dans lesquels $\hat{m}_1$ et $\hat{\mu}_2$ sont respectivement les moments d'ordre 1 et 2 estimés à partir des données.

– Si X est v.a. gamma de paramètres $(a > 0, b > 0, c)$ de distribution de probabilité

$$f_X(x, a, b, c) = \frac{1}{b} f_\gamma\left(\frac{x - c}{b}, a\right) = \frac{\exp\{-(x - c)/b\}(x - c)^{a-1}}{b^a \Gamma(a)}$$

alors, pour obtenir $\hat{a}$ et $\hat{b}$ il suffit de calculer

$$\hat{a} = \frac{(\hat{m}_1 + c)^2}{\hat{\mu}_2 - \hat{d}\{\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2\}} \qquad \text{et} \qquad \hat{b} = \frac{\hat{\mu}_2 - \hat{d}\{\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2\}}{(\hat{m}_1 + c)(1 - \hat{d})}$$

– Si X est une v.a. lognormale de paramètres $(a, b > 0, c)$ de distribution de probabilité

$$f_X(x, a, b, c) = \left(\frac{b}{x-c}\right) \phi\{a + b\ln(x-c)\}$$

alors pour obtenir $\hat{a}$ et $\hat{b}$ il suffit de calculer

$$\hat{a} = \frac{-\ln(\hat{v})}{\sqrt{2\ln(\hat{u})}} \qquad \text{et} \qquad \hat{b} = \frac{1}{\sqrt{2\ln(\hat{u})}}$$

avec

$$\begin{aligned} \hat{u} &= \frac{\sqrt{(1-\hat{d})(\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2)}}{\hat{m}_1 - \hat{c}} \\ \hat{v} &= \frac{(\hat{m}_1 - \hat{c})^2}{(1-\hat{d})\sqrt{(1-\hat{d})(\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2)}} \end{aligned}$$

– Si X est une v.a. de Weibull de paramètres $(a > 0, b > 0, c)$ de distribution de probabilité

$$f_X(x, a, b, c) = \frac{a}{b}\left(\frac{x-c}{b}\right)^{a-1} \exp\left\{-\left(\frac{x-c}{b}\right)^a\right\}$$

alors pour obtenir $\hat{a}$ il faut résoudre numériquement l'équation en a ci-dessous

$$\hat{\mu}_2 = (\hat{m}_1 - \hat{c})^2 \left\{\frac{\Gamma(2/a+1)}{(1-\hat{d})\Gamma^2(1/a+1)} - 1\right\}$$

et $\hat{b}$ s'obtient en calculant

$$\hat{b} = \frac{\hat{m}_1 - \hat{c}}{(1-\hat{d})\Gamma(1/\hat{a}+1)}$$

Une fois que les estimateurs initiaux $\hat{a}$, $\hat{b}$, $\hat{c}$ et $\hat{d}$ sont calculés, nous proposons d'utiliser un algorithme d'optimisation pour trouver les estimateurs $\hat{a}^*$, $\hat{b}^*$, $\hat{c}^*$ et $\hat{d}^*$ qui minimisent la distance de Kolmogorov $D = \max|F_Y(y, a, b, c, d) - \hat{F}(y)|$ où $\hat{F}(y)$ est la fonction de répartition empirique. La distance D la plus faible indique quelle distribution hybride doit être choisie pour modéliser les données.

La Figure 2 donne la fonction de répartition gamma
$F_X(x; 1.16822; 0.627758; 0; 0.875767)$ de la variable *"nombre de comptes SICAV"*. La valeur minimale de cette variable (*respectivement maximale*) est égale à 0 (*respectivement* 8).

Le taux de couverture des variables de la base pour la valeur minimale ($X = x_{min}$) est élevé à l'exception de la variable *"âge"*. Trois variables ont une valeur pour le taux de couverture inférieure à 20% (ces variables sont répertoriées dans la troisième catégorie et ont les valeurs suivantes : 12%, 18% et 19%), 10 variables ont un support compris entre 60% et 85% et pour finir, 27 variables ont un support supérieur à 85%. Nous pouvons vérifier l'importance de ne pas changer la valeur du paramètre c (*valeur minimale de* X) pour la fonction de répartition.

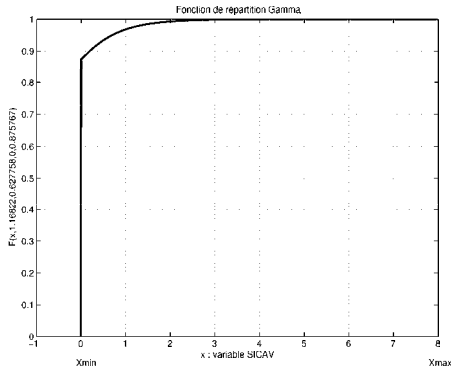

FIG. 2 – *Fonction de répartition gamma pour la variable "SICAV".*

4.4 Etude de l'Effet du Bruit sur les Mesures Quantitatives

Les expérimentations ont été faites en posant les valeurs suivantes pour les paramètres : $\gamma = 0,10$ (*pourcentage du bruit introduit dans la base*), $\alpha = 0,01$ (*valeur déterminant l'amplitude des intervalles pour les paramètres a, b et d*) et $k = 20$ (*nombre d'itérations*).

Ces paramètres ont été choisis de manière à injecter un bruit de nature réaliste dans les données étudiées. Il est raisonnable de considérer que les données peuvent contenir jusqu'à 10% de bruit ($\gamma = 0,10$), et que l'erreur associée à chaque valeur incorrecte n'excède pas 1% ($\alpha = 0,01$). Enfin, nous avons choisi d'observer l'effet moyen du bruit pour 20 itérations car les résultats ne semblent pas se modifier en augmentant le nombre d'itérations (*très faible valeur de la variance pour* 20 *itérations*).

Ces différents paramètres nous permettent de "contrôler" le bruit introduit dans les données. Nous pensons que le bruit obtenu est plus réaliste qu'un bruit uniforme ou gaussien car les données sont modifiées en fonction de leur distribution initiale.

L'axe des abscisses correspond au seuil d'élagage des associations, c-à-d la valeur minimale où dessous de laquelle l'association ne peut être jugée significative et par conséquent, retenue. L'axe des ordonnées correspond, pour la courbe en trait plein, au pourcentage de nouvelles règles qui sont apparues dans la base de données bruitée et pour la courbe en pointillé, au pourcentage de règles qui ont disparu des données bruitées.

La Figure 3 montre l'effet du bruit sur le coefficient de corrélation linéaire significatif (courbe de droite) et sur l'indice de vraisemblance du lien (courbe de gauche) et la Figure 4 montre l'effet du bruit sur l'intensité d'inclination.

Ayant introduit 10% de bruit dans les données, nous pensions observer l'apparition d'environ 10% de nouvelles règles (*incorrectes par définition*) , ainsi que la disparition de 10% de règles existantes. Les résultats observés pour ces deux mesures sont plutôt encourageants car, même pour des seuils d'élagage élevés, le bruit observé reste relativement inférieur au bruit introduit. De plus, la variance observée est proche de 0 ce qui renforce la qualité des résultats.

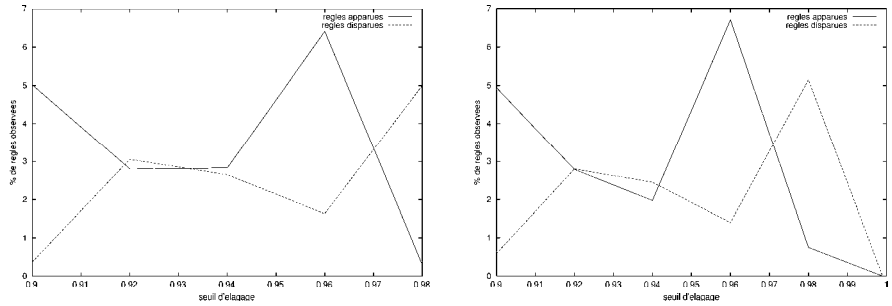

FIG. 3 – *Effet du bruit sur le coefficient de corrélation linéaire significatif (courbe de gauche) et l'indice de vraisemblance du lien (courbe de droite).*

Pour l'intensité d'inclination, les résultats obtenus sont nettement meilleurs que ceux observés pour les mesures précédentes car le pourcentage de règles nouvelles (obtenues par introduction du bruit dans les données) est très faible et proche de 0 (voir Figure 4).

Le bruit moyen observé pour la disparition des règles existantes est proche de 5%, à l'exception d'une valeur élevée pour un seuil d'élagage égal à 1. Cette augmentation du nombre de disparitions est liée au fait que peu de règles sont extraites lorsque le seuil d'élagage est fixé à 1. Ainsi, la disparition de peu de règles entraîne l'apparition de cette valeur "extrême" sur la courbe.

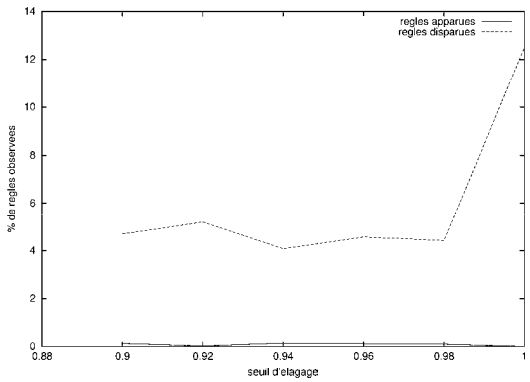

FIG. 4 – *Effet du bruit sur l'intensité d'inclination.*

5 Conclusion et perspectives

L'extraction non supervisée de connaissances dans des données volumineuses est particulièrement difficile car l'évaluation de la qualité des résultats obtenus repose es-

sentiellement sur les mesures de qualité utilisées pour obtenir ces résultats. Un des critères majeurs de qualité (selon nous), lorsque les données sont réelles et donc imparfaites, est de minimiser l'impact du bruit sur l'apparition de règles incorrectes. Les mesures ne vérifiant pas ce critère doivent être manipulées avec la plus grande précaution car les résultats fournis à l'expert, seul juge de la qualité des connaissances obtenues, peuvent s'avérer incomplets et surtout incorrects. Nos travaux ont permis de mettre en évidence, pour le corpus considéré et les mesures étudiées, que l'impact du bruit sur les connaissances obtenues n'est pas négligeable et varie en fonction des mesures de qualité utilisées.

Ces travaux doivent être poursuivis sur d'autres bases de données, de manière à valider les premiers résultats observés. La poursuite de ces travaux nécessite d'une part de pouvoir estimer efficacement les fonctions de distribution des variables des bases de données étudiées et d'autre part, de valider l'approche retenue pour introduire du bruit dans les données.

Références

Agrawal R., Imielinski T. et Swami A. N. (1993), Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp 207-216.

Agrawal R., Mannila H., Srikant R., Toivonen H. et Verkamo A. T. (1996), Fast discovery of association rules, Advances in Knowledge Discovery and Data Mining, AAAI Press, pp 307–328.

Azé J. et Kodratoff Y. (2002), Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association, Revue Extraction des connaissances et apprentissage, 1(4), pp 143-154.

Guillaume S. (2002), Discovery of Ordinal Association Rules, Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD02), pp 322–327.

Guillaume S. et Castagliola P. (2003), Étude de Mesures Quantitatives, Rapport Technique, Université de Blaise Pascal, Clermont-Ferrand II.

Mannila H., Toivonen H. et Verkamo A. I. (1994) , Efficient algorithms for discovering association rules, Workshop on Knowledge Discovery in Databases (KDD-94), AAAI Press, pp 181–192.

Lagrange J. B. (1997), Analyse Implicative d'un Ensemble de Variables Numériques ; Application au Traitement d'un Questionnaire à Réponses Modales Ordonnées, Prépublication 97-32 de l'Institut de Recherche Mathématiques de Rennes, Vol. XLVI(1), pp 1-27, 1997.

Larher A. (1991), Implication Statistique et Applications à l'Analyse de Démarches de Preuve Mathématique - Thèse d'État, Université de Rennes I.

Lerman I.C. (1981), Classification et analyse ordinale des données, Dunod.

Park J.S., Chan M-S. et Yu P.S. (1995), An Effective Hash Based Algorithm for Mining Association Rules, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pp. 175–186.

Pasquier N. (2000), Data Mining : Algorithmes d'extraction et de réduction des règles d'association dans les bases de données, Thèse, Universtié Blaise Pascal, Clermont-Ferrand II.

Saporta G. (1990), Probabilités, Analyse des Données et Statistique, Edition Technip.

Srikant R. et Agrawal R. (1996), Mining Quantitative Association Rules in Large Relational Tables, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, pp 1–12.

Summary

Knowledge extraction from real data is made difficult by the various defects of the data sets. The study of the effect of the noise in the data on the quality of the observed results improves our understanding of the quality measures. We present in this paper several quantitative measures discovering knowledge in data. For each of them, we perform an empiric survey of the effect of a realistic noise on financial data. Another quality criterion is derived from our analysis of the several measures we used. This new criterion, linked to the sensitivity to noise of the various measurements helps a better management of the choice of measures during the knowledge discovery process.