Extraction de « pépites » de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit

Jérôme Azé, Yves Kodratoff

CNRS, Laboratoire de Recherche en Informatique
Bât. 490, Université. Paris-Sud
91405 Orsay Cedex - France
{aze,yk}@lri.fr

Résumé. La plupart des méthodes permettant d'extraire des règles d'association dans les données sont basées sur l'utilisation de mesures et de seuils prédéfinis par l'expert pour optimiser la recherche des règles. Le choix des seuils permettant de séparer les règles intéressantes des règles triviales est difficile, même pour un expert du domaine étudié.

Si l'on considère que les données sont bruitées et que l'extraction d'informations du type « pépites » de connaissance (c'est-à-dire, règles ayant un faible support) peut intéresser l'expert, alors les méthodes classiques sont souvent mises en défaut dans de telles situations.

Nous proposons donc une nouvelle mesure d'extraction des règles d'association, appelée « moindre-contradiction ». Nous montrons que cette mesure (i) permet d'extraire des « pépites » de connaissance dans les données, sans pour autant être submergés par le nombre de règles ayant un faible support, (ii) se comporte légèrement mieux que les mesures classiques étudiées lorsque les données sont bruitées.

Mots Clés : règles d'association, mesures de qualité, bruit.

1 Introduction

La découverte non supervisée de règles d'association dans les bases de données est particulièrement intéressante et de nombreux travaux ont été effectués pour caractériser les motifs cachés dans les bases de données (Agrawal et al. 1993, Agrawal et Srikant 1995, Brin et al. 1997, Lavrač et al. 1999, Sahar 1999, Gras et al. 2001). Les effets du bruit et le comportement de ces algorithmes ont été peu étudiés, bien que les bases de données réelles soient rarement parfaites. Nous pensons qu'il est intéressant d'étudier le comportement des algorithmes d'extractions de règles d'association lorsque les données contiennent entre 1% et 10% de bruit. Il nous semble qu'au delà de 10% de bruit, les données sont trop imparfaites pour pouvoir être correctement analysées.

Ayant fait cette constatation, nous proposons, dans cet article, une mesure d'intérêt qui est moins sensible au bruit que la plupart des mesures classiques (que nous avons étudiées).
Comme nous l'avons déjà montré dans (Azé et Kodratoff 2002a, Azé et Kodratoff 2002b), de nombreuses mesures d'intérêt classiques, comme le support, la confiance

et la dépendance, sont très sensibles à différentes formes de bruits, et cette sensibilité est principalement liée au fait que ces mesures sont basées sur l'utilisation de seuils d'élagages permettant de déterminer un ensemble de règles intéressantes. L'ensemble des règles qui sont proches du seuil d'élagage peuvent basculer d'un côté ou de l'autre de ce seuil, lorsque les données sont bruitées, et provoquer ainsi l'apparition ou la disparition de nouvelles règles.

De plus, comme l'a déjà montré A. Freitas (Freitas 1998), les règles ayant un faible support peuvent être très intéressantes et il est difficile, même pour un expert du domaine, de définir un seuil d'élagage permettant de séparer efficacement, en deux classes, les règles intéressantes de celles qui ne le sont pas.

Une des approches possibles, est celle proposée par S. Sahar (Sahar 1999) qui présente les règles les plus générales à un expert. Si l'expert décide de classer une (ou plusieurs) de ces règles comme étant non intéressantes, alors cette règle et toutes ces spécialisations sont élaguées. Cette approche permet, à partir d'un ensemble de règles obtenues de manière « classique » (Agrawal et Srikant 1995), d'élaguer très rapidement l'ensemble des règles. Notre approche ne s'oppose pas à celle-ci, en revanche, nous essayons de réduire le plus possible l'ensemble initial de règles (qui peut être très volumineux).

Nous proposons donc un algorithme permettant d'extraire un ensemble de « pépites de connaissance » à partir d'une base de données, sans avoir besoin de demander à l'expert de définir un ensemble de seuils d'élagage. Ces pépites de connaissance sont représentées sous la forme de règles d'association, caractérisées par un faible support et une confiance strictement supérieure à 50%. La plupart des méthodes classiques produisent un volume de règles inexploitables par un expert dès que le support minimal devient trop faible. Ces méthodes ne permettent donc pas d'extraire efficacement les pépites de connaissance que nous recherchons.

Nous proposons donc une approche, basée sur la recherche non exhaustive des règles d'association les moins-contredites. Nous introduisons une mesure, appelée « moindre-contradiction » (Azé 2003), que nous noterons **contramin**. L'algorithme permettant d'extraire les pépites de connaissance recherchées est basé sur l'utilisation du contexte des règles en cours d'extraction. Parmi ces règles, seules les moins contredites sont retenues et proposées à l'expert. L'élagage ainsi effectué est donc dépendant des données et n'est pas confié directement à l'expert.

Pour illustrer l'intérêt des pépites de connaissance, considérons le problème classique de la recherche de connaissances dans des bases de données transactionnelles (ensemble de tickets de caisse d'un supermarché par exemple). Pour de telles données, l'expert (le directeur du supermarché) est intéressé par des règles nouvelles qui peuvent l'amener à faire des profits. Il semble raisonnable de considérer que les règles classiques et vérifiées par de nombreux consommateurs sont déjà connues et exploitées par les experts. Par exemple, les règles d'association *achat de beurre* $\rightarrow$ *achat de pain* ou *achat de lait* $\rightarrow$ *achat d'eau* semblent cohérentes et représentent des connaissances déjà connues du domaine. D'autres règles peuvent être très pertinentes mais n'être vérifiées que par une sous-population réduite des consommateurs, par exemple *achat de lait infantile et de couches pour bébé* $\rightarrow$ *achat de lingettes*.

Toutes ces connaissances sont déjà connues des experts et ceux-ci sont intéressés

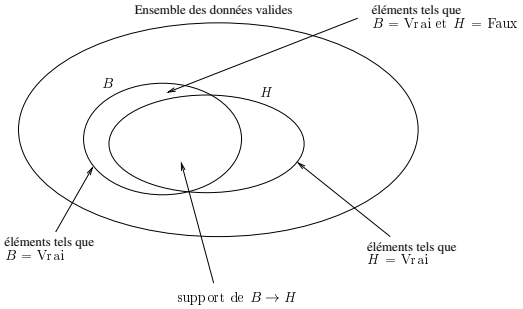

FIG. 1 – *Présentation de* $P(B \wedge H)$ *et de* $P(B \wedge \neg H)$.

par des connaissances nouvelles même si elles ne sont pas toujours vérifiées par une majorité de consommateurs. Nous pensons donc que les pépites de connaissance peuvent caractériser ces nouvelles connaissances tant recherchées. Ces pépites peuvent mettre en évidence de nouveaux comportements parmi les consommateurs et inciter les dirigeants à créer des promotions sur certains produits pour encourager la généralisation de ces nouveaux comportements. Une pépite potentiellement intéressante serait une règle de la forme $A \rightarrow B$ avec A un produit peu coûteux et B un produit onéreux. Une des exploitations possible de cette pépite étant : réaliser une promotion liant A et B, ceci dans le but d'inciter d'avantage de consommateurs à acheter le produit B, sachant que le produit A peut aussi les intéresser.

Nous présentons aussi différentes formes de bruits pouvant perturber les données. Nous proposons une amélioration de notre algorithme permettant de prendre en compte la nature bruitée des données et d'être donc moins sensible aux effets du bruit sur l'ensemble des règles d'association obtenu à partir de ces données.

Dans la section suivante, après avoir décrit les mesures les plus couramment utilisées, nous serons en mesure de définir plus précisément ce que nous entendons par « intérêt ».

2 Quelques mesures d'« intérêt »

L'ensemble des mesures présentées, support, confiance et dépendance, sont relatives à des variables discrètes ou booléennes. La Figure 1 présente les éléments de base permettant de comprendre le principe de ces différentes mesures. (Lavrač et al. 1999) présente un plus large ensemble de mesures de qualité. Il n'en reste pas moins que ces trois mesures constituent les éléments de base de la plupart des algorithmes existants.

Plaçons nous dans le cadre de la découverte de formes du type $B \rightarrow H$ (appelées « règles d'association entre deux éléments »). Nous utilisons la définition des règles d'association introduites dans (Agrawal et Srikant 1994).

Le **support** de $[B \rightarrow H]$ est défini par $P(B \wedge H)$. Il exprime la probabilité que B et H soient vrais ensemble. Notons que lorsque $P(B)$ et $P(H)$ sont tous les deux très importants (c'est-à-dire, ils déterminent chacun une colonne de la base ne conte-

nant pratiquement que des valeurs égales à $Vrai$ lorsqu'on se place dans un contexte booléen), l'étude de ce type de relation est sans réel intérêt bien que le support de ce type de relation soit toujours très élevé.

La **confiance** de la règle $[B \rightarrow H]$ est définie par $P(H|B) = \frac{P(B \wedge H)}{P(B)}$. Cette mesure exprime la probabilité conditionnelle que H soit vrai, sachant que B est vrai. Les règles ayant une confiance inférieure à 0.5 sont plus infirmées par les données plutôt que confirmées, ainsi ces règles ne sont pas « intéressantes ».

La **dépendance** de H par rapport à B est définie par $Abs(P(H|B) - P(H))$ si $P(B) \neq 0$, où Abs désigne la fonction valeur absolue (Kodratoff 2000). Si B n'est pas absurde, c'est-à-dire, si sa probabilité d'occurrence est non nulle, ce qui signifie que $P(H|B) = \frac{P(B \wedge H)}{P(B)}$ est calculable, alors, en présence de $B \rightarrow H$, la probabilité d'obtenir H sachant B doit être supérieure à $P(H)$. Ainsi, plus la différence est importante entre $P(H|B)$ et $P(H)$, plus la dépendance de la relation est élevée. Cependant, notons que lorsque $P(H)$ est élevé alors la dépendance est faible (si $P(H)$ est élevé, proche de 1, nous avons $P(H|B) \approx P(H)$ et donc $P(H|B) - P(H) \approx 0$).

Nous allons illustrer, au moyen des trois cas présentés dans la Figure 2, certaines propriétés de la dépendance. Les trois cas présentés sur la Figure 2 représentent un cas particulier et intéressant de dépendance. Dans chacun de ces cas, nous avons $P(H|B) = 1$, mais les raisonnements qui suivent s'étendent trivialement au cas où $P(H|B)$ est « grand ».

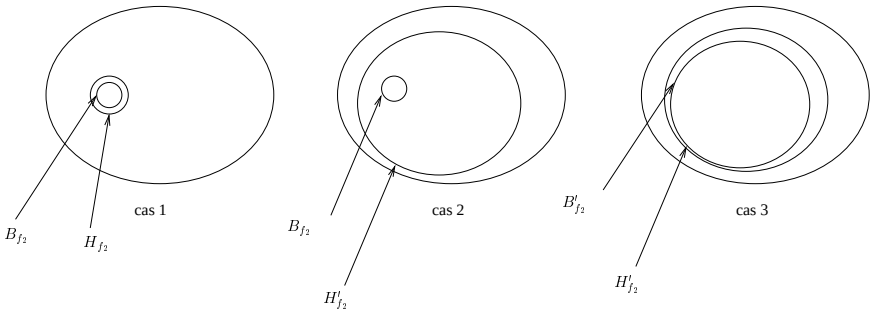

FIG. 2 – *Trois cas illustrant différents comportements de la dépendance.*

Dans le cas 1, $P(H_{f2})$ est faible, ainsi ce cas correspond au cas d'une dépendance élevée, combinée à un très faible support. Ce cas met en évidence des relations présentant un intérêt très élevé car les ensembles considérés sont très petits (ils peuvent donc être inconnus de l'expert), et les relations considérées ne sont jamais infirmées.

Dans le cas 2, B_{f2} est « noyé » dans H'_{f2} et, il est relativement intuitif de considérer que H'_{f2} dépend moins de B_{f2}, alors que H_{f2} dépend fortement de B_{f2}, dans le cas 1. Dans ce cas, la dépendance $1 - P(H'_{f2})$, est faible, comme nous nous y attendions.

Le comportement de la dépendance sur ces deux premiers cas est conforme à nos attentes. Pour ces deux cas, l'utilisation de la dépendance, comme mesure d'intérêt, peut sembler acceptable.

L'étude du cas 3, pour lequel $P(H'_{f2})$ et $P(B'_{f2})$ sont tous les deux très élevés, montre que la dépendance est faible. La règle d'association $B'_{f2} \rightarrow H'_{f2}$ correspond au cas d'une relation relativement triviale et sans réel intérêt, en terme de connaissances nouvelles. Cependant, nous allons voir que la définition de la dépendance est relativement contre-intuitive dans ce cas. Nous pouvons noter une contradiction entre l'intuition classique de la dépendance et ce que nous appellerons la « dépendance discrète » (c'est-à-dire où les attributs sont de nature booléenne). L'intuition classique, utilisée par de nombreux auteurs, repose sur la théorie des probabilités qui précise que la probabilité conjointe de deux événements indépendants, $P(B \wedge H)$, est égale à $P(B) * P(H)$. Ainsi, selon l'intuition classique, la dépendance entre deux évènements, B et H, est élevée lorsque la différence entre $P(B \wedge H)$ et $P(B) * P(H)$ est importante.

Dans le cas discret que nous étudions ici, B'_{f2} et H'_{f2} sont manifestement très dépendants car $B'_{f2} = Vrai$ nous permet de prédire $H'_{f2} = Vrai$ avec une confiance donnée par $\frac{P(B'_{f2} \wedge H'_{f2})}{P(B'_{f2})}$. Cependant, nous avons bien $P(H'_{f2}) * P(B'_{f2}) \approx 1 \approx P(B'_{f2} \wedge H'_{f2})$.

L'objectif de cet article n'est pas d'étudier les détails de ce paradoxe, mais nous voulons le mettre en évidence de manière à montrer que la mesure que nous proposons, la **moindre-contradiction (contramin)**, présente, comme la dépendance, des propriétés paradoxales, lorsque le cas 3 est pris en considération.

Cette mesure est construite de manière à favoriser les cas où B est pratiquement totalement inclus dans H, c'est-à-dire où $[P(B \wedge H) - P(B \wedge \neg H)]$ est « étonnamment » élevé (c'est-à-dire, des cas où l'implication $B \rightarrow H$ est rarement infirmée). Pour prendre en considération le cas 2 présenté ci-dessus, nous proposons de normaliser notre mesure par rapport à $P(H)$. La contramin, d'une règle $B \rightarrow H$, est donc définie de la manière suivante :

$$contramin(B \rightarrow H) = \frac{P(B \wedge H) - P(B \wedge \neg H)}{P(H)}$$

Le comportement de cette mesure est satisfaisant dans les cas 1 et 2. Nous avons bien $contramin(B_{f2} \rightarrow H_{f2}) > contramin(B_{f2} \rightarrow H'_{f2})$. Par contre, dans les cas où $\frac{P(B_{f2})}{P(H_{f2})} \approx \frac{P(B'_{f2})}{P(H'_{f2})}$, alors notre mesure ne permet pas de différencier le cas 1 du cas 3. Ce problème doit être résolu en utilisant d'autres mesures (par exemple le support de la règle).

Le support, la confiance et la dépendance présentent différents aspects de l'intérêt d'une règle. La définition de la contramin rassemble quelques « qualités » de chacune de ces mesures. La contramin nous permet d'extraire et d'ordonner des règles intéressantes à partir d'une base de données.

Précisons les différents aspects de la notion d'intérêt que nous allons utiliser dans la suite de cet article : (i) une règle intéressante doit être plus confirmée par les données qu'infirmée (sinon, elle ne serait pas « sûre »), (ii) elle ne doit pas être vraie pour toutes les données (sinon elle serait triviale), (iii) la dépendance entre la prémisse de la règle, B, et la conclusion de celle-ci, H, doit être élevée (sinon, nous aurions une règle

« faible »), (iv) elle doit permettre d'améliorer la connaissance d'un expert du domaine, (v) elle doit être peu sensible au bruit présent dans les données. Nous n'étudierons pas la condition (iv) dans la suite de cet article. La condition (v) sera abordée en détail dans la section 5.
Pour satisfaire les conditions (i) et (ii), nous ne retenons que les règles $B \rightarrow H$ ayant une confiance supérieure à $0,5$.
Comme nous l'avons indiqué précédemment, la condition (iii) n'est pas remplie par la dépendance, car comme nous pouvons le voir dans la Figure 2, pour les cas 2 et 3, la dépendance entre B'_{f2} et H'_{f2} est faible alors que B'_{f2} permet de prédire H'_{f2} de manière parfaite dans les deux cas. C'est pourquoi, nous allons étudier contramin définie précédemment.

3 Algorithme pour la découverte des règles les moins-contradictoires

Notre objectif principal étant lié à la découverte de « pépites » de connaissance dans les données, nous ne pouvons pas fixer de seuil minimal pour le support. De plus, contramin ne possède pas de propriété de monotonie ou d'anti-monotonie (une mesure $\mathcal{S}$ est anti-monotone si $(\forall R, R'\ tq\ R \in R'\ alors S(R) > S(R'))$. Il est donc difficile (voire impossible) de rechercher de manière exhaustive l'intégralité des règles d'association les moins-contradictoires.
C'est pourquoi, nous avons été conduits à limiter le problème que nous traitons : nous ne recherchons que des règles d'association ayant les propriétés suivantes : (i) être les règles présentant les plus grandes valeurs pour contramin, (ii) être telles que les prémisses des règles ne contiennent pas plus de K attributs et telles que la conclusion de celles-ci soit réduite à un seul attribut.

Une extension limitée à des règles ayant plusieurs conclusions est en cours, les limites de l'extension seront dictées par un expert du domaine.

Nous proposons un algorithme permettant d'extraire les règles d'association les moins-contredites satisfaisant les points (i) et (ii).

Pour satisfaire la première condition, (i), nous proposons d'utiliser une approche classique en Analyse de Données (Daudé 1992), à savoir normaliser les valeurs de chaque mesure par rapport à l'ensemble des mesures observées. Cette normalisation est effectuée de la manière suivante : (1) la valeur de la mesure est centrée par rapport à la valeur moyenne des mesures observées, puis (2) cette valeur est réduite par rapport à l'écart-type observé pour ces mêmes règles.

La deuxième condition, (ii), est justifiée par le fait que nous considérons que des règles d'association ayant un grand nombre d'attributs en prémisse et plus d'un attribut en conclusion, sont difficilement interprétables. Même pour un expert du domaine, de trop longues règles sont difficilement compréhensibles, c'est pourquoi, nous avons choisi de limiter le nombre d'attributs en prémisse et en conclusion des règles recherchées.

Le Tableau 1 présente les notations utilisées dans l'algorithme 1.

Algorithme 1 ExtrairePépites($\mathcal{D}_n^p, K_{max}, min_{sup}$)

Entrée:

$\mathcal{D}_n^p$: la base de données étudiée contenant n individus décrit par p attributs

K_{max} : nombre maximal d'attributs en prémisse, défini par l'utilisateur

min_{sup} : support minimal pour les pépites de connaissance

Sortie:

$\mathcal{E}$: ensemble des règles d'association les moins-contredites par les données contenant au plus K_{max} attributs en prémisse et telles que $\{confiance(R) > 0.5, \forall R \in \mathcal{E}\}$.

Début

$K = 1$

$T_K = \mu_0 = \sigma_0 = 0$

$\mathcal{E} = \emptyset$

$\mathcal{E}_1 = \{$ *l'ensemble de tous les attributs de la base étudiée* $\}$

Pour tout $(X_j \in \mathcal{E}_1)$ **faire**

tant que $(\mathcal{E}_K \neq \emptyset)$ and $(T_K < 1)$ and $(K \leq K_{max})$ **faire**

-- *Génération de* $\mathcal{C}_K$ *à partir de* $\mathcal{E}_K$

si $(K = 1)$ **alors**

-- *Cas particulier où* $\mathcal{E}_K$ *ne contient que des attributs*

$\mathcal{C}_K = \{X_i \rightarrow X_j / X_i \in \mathcal{E}_1 \ , \ X_j \in \mathcal{E}_1 \ , \ i \neq j \ , \ support(X_i \rightarrow X_j) > min_{sup}\}$

sinon

-- *Cas général où* $\mathcal{E}_K$ *contient des règles*

$\mathcal{C}_K = \{X_i, X_l \rightarrow X_j / X_i \rightarrow X_j \in \mathcal{E}_K \ , \ X_l \rightarrow X_j \in \mathcal{E}_K \ , \ i \neq l \ , \ support(X_i, X_l \rightarrow X_j) > min_{sup}\}$

fin si

$\mathcal{E}_K^+ = \{R \in \mathcal{C}_K / contramin(R) > T_K\}$

$\mathcal{E}_K^- = \{R \in \mathcal{C}_K / contramin(R) \leq T_K\}$

si $(\mathcal{E}_K^+ = \emptyset)$ **alors**

$T_{K+1} = T_K$ -- *Dans ce cas,* μ_K *et* σ_K *ne sont pas calculables*

$\mathcal{E}_{K+1} = \mathcal{E}_K^-$

sinon

Calcul de μ_K et σ_K

$\mathcal{E}'_K = \{R \in \mathcal{E}_K^+ / \frac{contramin(R) - \mu_K}{\sigma_K} > 1\}$

$\mathcal{E} = \mathcal{E} \cup \mathcal{E}'_K$

-- *Préparation du niveau suivant*

$\mathcal{E}_{K+1} = \mathcal{E}_K^- \cup (\mathcal{E}_K^+ - \mathcal{E}'_K)$ -- $\{$*règles non proposées à l'expert*$\}$

$T_{K+1} = \mu_K + \sigma_K$

fin si

$K = K + 1$

fin tant que

fin pour

Retourner $\mathcal{E}$

Fin

symbole	signification
$\mathcal{D}_n^p$	base de données contenant n individus décrits par p attributs
K	nombre d'attributs dans la prémisse d'une règle d'association $A \rightarrow B$
$\mathcal{E}_K$	ensemble de toutes les règles d'association utilisé comme ensemble générateur pour $\mathcal{C}_K$
$\mathcal{C}_K$	ensemble des règles d'association « candidates » obtenues à partir de $\mathcal{E}_K$ contenant les règles d'association les moins-contradictoires
T_K	seuil d'élagage pour les règles d'association les moins-contradictoires
$\mathcal{E}_K^+$	sous-ensemble des règles de $\mathcal{C}_K$ dont la moindre contradiction se situe au delà du seuil T_K
$\mathcal{E}_K^-$	sous-ensemble des règles de $\mathcal{C}_K$ dont la moindre contradiction se situe en dessous du seuil T_K
μ_K	moyenne des règles d'association les moins-contradictoires appartenant à $\mathcal{E}_K^+$
σ_K	écart-type des règles d'association les moins-contradictoires appartenant à $\mathcal{E}_K^+$
$\mathcal{E}'_K$	ensemble des règles d'association les moins-contradictoires parmi $\mathcal{E}_K^+$
$\mathcal{E}$	ensemble des règles d'association telles que la moindre-contradiction soit la plus élevée

TAB. 1 – *Notations utilisées dans l'algorithme 1.*

3.1 Détails de l'étape itérative : $K \geq 1$ (c'est-à-dire, B contient K attributs)

Les règles appartenant à $\mathcal{E}'_K$ sont conservées sans être modifiées car, à l'étape $K+1$, les nouvelles règles obtenues ne doivent pas être des spécialisations de règles déjà obtenues à l'étape K. Cette heuristique est définie de manière à ne pas surcharger l'expert. Notre mesure ne possédant pas de propriété de monotonie, nous ne pouvons pas garantir qu'il n'existe pas de règle $B \wedge X \rightarrow H$ plus spécifique que $B \rightarrow H$ et qui soit moins contredite. Cependant, si nous choisissons de spécialiser l'intégralité des règles les moins contredites, nous obtenons alors un ensemble de règles trop volumineux pour pouvoir être soumis à l'expert. Nous pensons qu'il est plus aisé pour l'expert d'examiner un ensemble de règles générales et d'étudier, si besoin, un sous-ensemble de règles issu de la spécialisation d'une règle qu'il aura explicitement choisie.

Ainsi, cette heuristique peut être considérée comme une implantation des travaux de S. Sahar (Sahar 1999).

Il est important de noter que l'utilisation de cette heuristique implique que les règles soient considérées comme des quantifications universelles. Donc l'expert peut rejeter une généralisation et accepter une de ces spécialisations (par exemple, *tous les humains aiment un autre humain* : rejetée et *Pierre Curie aimait Marie* : acceptée). L'expert doit donc avoir conscience de ce comportement et agir en conséquence lorsqu'il analyse les règles.

Comme le nombre de règles situées au-dessus d'un seuil de contramin prédéfini augmente très rapidement avec K, et comme nous voulons éviter d'engendrer trop de règles, nous devons définir une nouvelle heuristique permettant de faire en sorte que le seuil d'élagage augmente avec K. Nous nous focalisons alors sur la spécialisation des règles appartenant à $\mathcal{E}_{K+1} = \mathcal{E}_K^- \cup (\mathcal{E}_K^+ - \mathcal{E}'_K)$ et nous initialisons le nouveau seuil

d'élagage, pour les règles les moins-contradictoires, avec $\mu_K + \sigma_K$, de manière à ne retenir que les « meilleures » règles les moins-contradictoires.
Il est aisé de montrer que cet algorithme converge lorsque K augmente. A chaque itération de l'algorithme, les règles obtenues, appartenant à $\mathcal{E}'_K$, sont telles que les valeurs de contramin sont strictement supérieures à celles obtenues à l'étape précédente. Ainsi, si les conditions d'arrêt $\mathcal{E}_K = \emptyset$ ou $K > K_{max}$ ne sont pas atteintes, nous pouvons garantir que la condition d'arrêt $T_K = 1$ sera atteinte. En effet, par construction, nous avons $\forall K, \mu_K > T_K$, $\sigma_K \geq 0$ et $T_{K+1} = \mu_K + \sigma_K$, donc $\mu_K > \mu_{K-1}$ et $T_{K+1} > T_K$. Nous avons $T_0 = 0$, T_K est strictement croissant en fonction de K, donc, $\exists K \ tq \ T_K = 1$.

3.2 Estimation de la complexité de l'algorithme

Dans le pire des cas, l'ensemble $\mathcal{E}_K^+$ est vide $\forall K, 1 \leq K \leq K_{max}$ et le support des règles n'est jamais nul. Dans ce cas, l'algorithme ne trouve aucune pépite et aucun élagage ne peut être réalisé à l'étape K pour réduire le nombre de règles candidates à l'étape $K+1$. Le coût de l'algorithme est donc exponentiel en fonction du nombre p d'attributs de la base de données car, pour chaque valeur de K, le nombre de règles candidates est égal à C_p^K.

Dans le cas général, pour chaque valeur de K, l'ensemble $\mathcal{E}_K^+$ contient des règles qui pourront être utilisées pour élaguer l'espace de recherche et ainsi minimiser le coût de calcul des règles à l'étape $K+1$. L'estimation exacte du gain représenté par cet élagage est difficile à réaliser. Si on note k le premier niveau sur lequel l'élagage est effectué et N_K^+ le nombre de règles élaguées, le gain de notre algorithme, par rapport à Apriori , est égal $C_p^{k+1} - G_p^k$, où G_p^k représente le nombre de règles non spécialisées car élaguées au niveau k. Or le calcul exact de G_p^k est difficile à cause des recouvrements entre règles élaguées. Par exemple, considérons les attributs $\{A, B, C, D, E, F\}$; si les règles $A, B \rightarrow C$ et $A, D \rightarrow C$ appartiennent à $\mathcal{E}_2^+$ alors les règles $A, B, D \rightarrow C$, $A, B, E \rightarrow C$ et $A, B, F \rightarrow C$ sont élaguées par la règle $A, B \rightarrow C$ et les règles $A, D, B \rightarrow C$, $A, D, E \rightarrow C$ et $A, D, F \rightarrow C$ sont élaguées par la règle $A, D \rightarrow C$. Déterminer le recouvrement de ces deux ensembles de règles, ici $A, B, D \rightarrow C$, est un problème non trivial.

Le coût de l'approche reste exponentiel mais inférieur à celui d'Apriori .

4 Les bases de données étudiées

Avant d'utiliser notre méthode sur des bases de données réelles, nous avons choisi de la tester sur huit bases de données de l'UCI (Blake et Merz 1998). Les bases de l'UCI sont initialement prévues pour évaluer des systèmes d'apprentissage supervisés. Notre approche étant non supervisée, nous assimilons les classes contenues dans ces bases à de simples attributs. Ainsi, les conclusions des règles d'association obtenues ne sont pas limitées aux classes définies dans les données.

Dans cette première série d'expériences, nous n'avons retenu que des bases de données discrètes, puis nous avons transformé toutes ces bases en bases de données booléennes.

Base	# d'attributs discrets	# d'attributs booléens	# d'enregistrements	Valeurs absentes?
car	7	25	1728	non
monks-1	7	19	432	non
monks-2	7	19	432	non
monks-3	7	19	432	non
mushrooms	23	125	8124	oui
nursery	9	32	12960	non
tic-tac-toe	10	29	958	non
votes	17	49	435	oui

TAB. 2 – *Quelques bases de données étudiées.*

Dans la base « mushrooms », les valeurs manquantes correspondent à une mesure non effectuée. Nous les remplaçons par la valeur « faux » ce qui présente le seul défaut d'augmenter artificiellement la contradiction des règles. Dans la base « votes », les valeurs manquantes correspondent à un vote non exprimé. Elles ont été remplacées par la valeur « non exprimé » qui s'ajoute aux valeurs discrètes possibles.

4.1 Résultats obtenus

Lors de la validation de notre approche, nous avons voulu mettre en évidence les deux points suivants :

1. montrer que le volume de règles engendrées par notre approche est inférieur au volume de règles engendrées par APRIORI, avec les mêmes contraintes initiales : $support > 0$ et $confiance > 0,5$;

2. montrer que l'ensemble des règles obtenues par notre approche représente une source de connaissances intéressantes pour l'expert du domaine ;

Lors de la première phase de la validation, le volume de règles extraites par notre algorithme est comparé avec les règles extraites par l'algorithme APRIORI (Agrawal et al. 1993). Les contraintes suivantes sont utilisées pour effectuer l'extraction : $support_{mini} = 0$, $confiance_{mini} = 0,5$, $K_{max} = 3$.

Les résultats obtenus sont présentés dans le tableau 3.

Nous pouvons voir que le volume de règles engendré par notre approche est nettement inférieur à celui obtenu par une approche de recherche exhaustive des règles d'association. L'application d'un post-traitement, reproduisant le comportement de notre algorithme, sur l'ensemble des règles engendrées par APRIORI, en éliminant toutes les règles de la forme $X \wedge Z \rightarrow Y$ si la règle $X \rightarrow Y$ est telle que la moindre contradiction de cette règle est supérieure au seuil déterminé par la moyenne et l'écart-type observé sur les règles du même type. L'un des avantages de notre approche réside dans le fait que nous ne sommes pas contraints de produire l'intégralité des règles pour ensuite devoir les élaguer à l'aide d'heuristiques. Ainsi, les spécialisations des règles « moindre-contradictoires » ne sont pas engendrées.

Les résultats obtenus sont détaillés dans le tableau 4.

Base	APRIORI	notre approche
car	1617	5
monks-1	1935	17
monks-2	1796	20
monks-3	1987	17
mushrooms	680088	386
nursery	3319	7
tic-tac-toe	8094	32
votes	17681	232

TAB. 3 – *Taille des ensembles de règles obtenues.*

Bases	APRIORI			notre approche		
	$K=1$	$K=2$	$K=3$	$K=1$	$K=2$	$K=3$
car	30	281	1306	2	3	0
monks-1	46	366	1523	9	8	0
monks-2	50	343	1403	10	10	0
monks-3	46	372	1569	9	8	0
mushrooms	1727	48316	630045	265	121	0
nursery	22	378	2919	4	3	0
tic-tac-toe	28	802	7264	0	32	0
votes	628	14338	161925	114	118	0

TAB. 4 – *Détail des ensembles de règles obtenus.*

La deuxième phase de la validation nécessite l'expertise d'un spécialiste du domaine. Il est indispensable de faire valider les ensembles de règles obtenus par un expert, de manière à valider l'approche globale. Le faible nombre de règles obtenues n'est pas garant de la qualité de celles-ci, seul un expert peu apprécier la qualité des règles.

Malheureusement, les experts sont rares et nous ne disposons pas de spécialistes pour les bases de données étudiées. Cependant, en utilisant nos connaissances généralistes, nous avons pu valider les règles obtenues à partir des bases *car*, *nursery* et *tic-tac-toe*. Ces règles, peu nombreuses et facilement interprétables pour un néophyte, représentent des connaissances valides (pour *car* et *nursery*) mais non nouvelles, compte tenu de nos connaissances. Notre manque d'expertise sur ces données ne nous permet pas de conclure sur la qualité des règles obtenues par rapport aux connaissances « attendues » à partir de ces bases. Les règles obtenues à partir de *tic-tac-toe* ne sont pas très informatives.

Ces résultats ne permettent pas d'évaluer la capacité de notre algorithme à extraire les pépites de connaissances. Seuls des résultats validés par des experts peuvent être considérés comme une validation positive pour notre approche.

Nous avons étudié des données issues de processus ancrés dans la vie réelle et intéressant des expert. Ces données ont été obtenues à partir de corpus de textes et sont relatives, d'une part à des introductions d'articles scientifiques du domaine de la fouille de données écrits en langue anglaise et d'autre part à des textes résumant des questionnaires de ressources humaines écrits en langue française et fourni par la société PerformanSe[1]. Les résultats obtenus sur ces deux bases de données correspondent à des règles d'association extraites en utilisant la moindre contradiction. Les experts ont analysé et validé les règles obtenues. Ces règles ont été considérées comme intéressantes et une partie d'entre elles représentaient des connaissances nouvelles pour les experts (Kodratoff et al. 2003).

Ces deux expérimentations sur des données réelles nous ont donc permis de valider notre approche.

La section suivante présente une étude comparative du comportement de la moindre contradiction avec d'autres mesures de qualité en présence de données bruitées. Nous verrons qu'il est difficile de travailler avec des données bruitées et que la moindre contradiction est la mesure qui présente le meilleur comportement compte tenu du cadre expérimental étudié.

5 Étude du bruit

Comme nous étudions le problème de la détection de règles d'association, tous les attributs peuvent se trouver soit dans la prémisse, soit dans la conclusion d'une règle. Ainsi, l'ensemble des attributs peut être affecté par le bruit, mais un attribut bruité ne peut modifier que les règles contenant cet attribut donné. C'est pourquoi nous nous sommes focalisés sur l'étude de trois différents types de bruit.

a. La première méthode consiste à introduire du bruit dans un attribut X, par exemple,

1. http://www.perfomanse.fr

et à étudier les effets du bruit sur les règles contenant cet attribut donné. L'ensemble des règles ne contenant pas l'attribut X n'est pas modifié et l'effet du bruit est lié au nombre de règles contenant l'attribut X. De manière à préciser notre approche, considérons le cas d'une règle liant deux attributs binaires, par exemple $X = Vrai$ et $Y = Vrai$ (ce type de règle sera celui que nous étudierons dans la suite de cette section). Supposons que cet attribut X soit bruité, avec 5% de bruit. Nous introduisons le bruit en renversant 5% des valeurs de l'attribut X, c'est-à-dire en changeant les valeurs Vrai par Faux et inversement. La quantité de règles telles que $X = Vrai$ et $Y = vrai$ est alors modifiée.

En fait, nous observons deux changements dus à ce type de bruit :

- Quand une valeur $Vrai$ devient égale à $Faux$, des règles contenant $X = Vrai$ et $Y = Vrai$ vont disparaître.
- Quand une valeur $Faux$ devient égale à $Vrai$, des règles contenant $X = Vrai$ et $Y = Vrai$ vont apparaître.

Avec cette première méthode, nous pouvons isoler le bruit et comprendre ses effets. La proportion de règles qui disparaissent est liée à la présence de règles contenant $X = Vrai$ et $Y = Vrai$ dans la base de données ; inversement, la proportion de règles créées par le bruit est liée à la présence de règles contenant $X = Faux$ et $Y = Vrai$ dans la base de données.

L'algorithme 2 permet d'introduire ce type de bruit dans une base de données.

Pour faciliter la compréhension des algorithmes, nous avons introduit les notations suivantes. Soit $\mathcal{D}_n^p$ une base de données décrites par un ensemble $\mathcal{O} = \{o_i | 1 \leq i \leq n\}$ d'objets et un ensemble $\mathcal{A} = \{a^j | 1 \leq j \leq p\}$ d'attributs booléens. Nous notons α_i^j la valeur de l'attribut a^j pour l'objet o_i. Pour éviter toute ambigüité, nous notons $\mathcal{D}_n^p(\alpha_i^j)$ la valeur de l'attribut a^j pour l'objet o_i dans la base de données $\mathcal{D}_n^p$. Lorsqu'il n'y a pas d'ambigüité sur les notations, n et p peuvent être omis.

b. La seconde méthode consiste à introduire le bruit de manière aléatoire dans la base de données. Le but de cette expérience est de montrer comment les différentes mesures réagissent lorsque les attributs sont bruités de manière aléatoire avec un faible pourcentage de bruit.

L'algorithme 3 permet d'introduire ce type de bruit dans les données.

c. La troisième méthode consiste à introduire différents niveaux de bruit sur quelques attributs de la base. Le but de cette expérience est d'étudier la sensibilité au bruit des mesures lorsque les données sont globalement peu bruitées (1%) mais que ce bruit est lié à quelques attributs seulement. Nous avons l'impression que ce type de bruit reflète mieux les situations réelles auxquelles nous sommes confrontés. En effet, il est peu probable que toutes les données soient bruitées, c'est-à-dire incorrectes. Si tel était le cas, *a priori* aucune méthode ne pourrait extraire des connaissances valides à partir de ces données. Par contre, il est relativement

Algorithme 2 IntroduireBruit-a

Entrée: $\mathcal{D}_n^p$: base de données à bruiter
Entrée: j : indice de l'attribut bruité (entre 0 et 1)
Entrée: α_{noise} : bruit introduit dans les données
Sortie: $\mathcal{D}'$: base de données bruitée

```
Début
nb_objets_a_bruiter = int(α_noise * p)        -- int(x) : partie entière de (x)
D' ← D
pour i = 1; i ≤ n; i++ faire
    nonBruite[i] = 1
fin pour
pour (k = 1; k ≤ nb_objets_a_bruiter; k++) faire
    Répéter
        i ← entier choisi aléatoirement entre 1 et n
    Jusqu'à ce que (nonBruite[i] = 1)
    nonBruite[i] = 0
    D'(α_i^j) = not D(α_i^j)        -- introduction du bruit en inversant la valeur de D(α_i^j)
fin pour
Retourner D'
Fin
```

Algorithme 3 IntroduireBruit-b

Entrée: $\mathcal{D}_n^p$: base de données à bruiter
Entrée: α_{noise} : bruit introduit dans les données (entre 0 et 1)
Sortie: $\mathcal{D}'$: base de données bruitée

```
Début
nb_objets_a_bruiter = int(α_noise * p * n)        -- int(x) : partie entière de (x)
D' ← D
pour (i = 1; i ≤ n; i++) faire
    pour (j = 1; j ≤ p; j++) faire
        nonBruite[i][j] = 1
    fin pour
fin pour
pour (k = 1; k ≤ nb_objets_a_bruiter; k++) faire
    Répéter
        i ← entier choisi aléatoirement entre 1 et n
        j ← entier choisi aléatoirement entre 1 et p
    Jusqu'à ce que (nonBruite[i][j] = 1)
    nonBruite[i][j] = 0
    D'(α_i^j) = not D(α_i^j)        -- introduction du bruit en inversant la valeur de D(α_i^j)
fin pour
Retourner D'
Fin
```

raisonnable de considérer que la majorité des attributs de la base sont fiables, et que seulement quelques attributs sont bruités.

Par exemple, considérons une base d'individus pour lesquels, les informations suivantes sont renseignées: âge, taille, poids, sexe, nationalité. Cette base de données est remplie par un employé de mairie disposant de la carte d'identité de chaque individu. La probabilité pour que les informations: taille, sexe, âge et nationalité soient incorrectes est très faible car ces informations sont présentes sur une carte d'identité. Par contre, le poids n'y figure pas, la probabilité d'erreur est donc beaucoup plus élevée pour cet attribut que pour les quatre précédents.

Cette troisième méthode essaye donc de rendre compte de ce phénomène en introduisant différents niveaux de bruit, sur quelques attributs de la base. Le nombre d'attributs à bruiter est un paramètre contrôlé par l'expert du domaine étudié. En fonction du type de données manipulées et des conditions dans lesquelles ces données ont été collectées, le nombre d'attributs bruités peut varier significativement.

De manière à mieux comprendre l'impact de ce bruit sur les règles obtenues, nous devons introduire du bruit sur quelques attributs, par exemple 3, et ce pour différents niveaux de bruit, par exemple 1%, 5% et 10%. Pour chaque triplet de couple $(Attribut, bruit)$, nous observons un certain bruit résultant dans les règles obtenues. La moyenne de ces différentes valeurs de bruit nous permet d'apprécier l'impact moyen de ce bruit sur notre approche. La combinatoire de cette méthode est élevée et l'évaluation de ce bruit sur une base de données telle que « Mushrooms » est très coûteuse en temps de calcul.

Dans (Azé et Kodratoff 2002a, Azé et Kodratoff 2002b), nous avons déjà étudié le comportement de la moindre-contradiction avec ces trois différents types de bruit pour la détection des règles d'association contenant exactement un attribut en prémisse et en conclusion. Nous avons observé que les deux dernières formes de bruit sont celles qui ont les effets les plus importants sur les règles les moins-contradictoires.
Les expériences précédentes ont montré que la seconde forme de bruit est celle qui dégrade le plus l'ensemble des règles obtenues. Nous avons donc choisi d'améliorer notre algorithme en éliminant les règles d'association qui sont trop sensibles à cette seconde forme de bruit.

En fait, nous proposons de modifier notre algorithme de manière à pouvoir rejeter les règles d'association qui présentent une très grande valeur de la moindre-contradiction, proche de un, et qui sont très sensibles au bruit. Ces règles d'association ne sont jamais (ou rarement) infirmées par les données mais elles possèdent un très faible support. Dès que nous introduisons une faible quantité de bruit, ce type de règles d'association disparaît de l'ensemble des règles les moins-contradictoires, ceci étant dû à l'apparition de contradictions liées au bruit. Ces cas correspondent aux « pépites » de connaissances recherchées, mais, lorsque les données sont bruitées, ces « pépites » sont particulièrement instables. Cependant, il existe un support minimal tel que ces relations restent stables même en présence de bruit. Nous nous proposons donc d'essayer de déterminer de manière automatique ce support minimal.

Weiss et Hirsh ont montré dans (Weiss et Hirsh 1998) que les « small disjuncts » sont très sensibles au bruit et qu'ils dégradent significativement les résultats du processus d'apprentissage. Ces observations, effectuées dans le cadre de l'apprentissage supervisé, peuvent être réutilisées dans le cadre de la découverte non supervisée de règles d'association, où les règles d'association ayant un très faible support peuvent être comparées aux « small disjuncts » de par le fait que seuls peu d'individus de la base de données vérifient ces règles.
Pour prendre ce problème en considération, nous introduisons le concept de « support minimum pour la résistance à un α-bruit ». Ce concept est évalué comme étant le support stable pour lequel les règles d'association, ayant une valeur élevée (proche de 1) pour la moindre-contradiction, ne disparaissent plus lorsque nous introduisons N fois $\alpha\%$ de bruit dans les données. Ce support minimum est utilisé comme seuil d'élagage pour l'étape suivante de l'algorithme, et un support est considéré comme stable lorsqu'il ne varie plus lors des différentes itérations.

Il existe en effet des règles d'association ayant une moindre contradiction très élevée mais un support très faible. Ces règles correspondent, par définition, aux pépites de connaissance que nous recherchons. Le problème est qu'il est difficile de déterminer, lorsque leur support est très faible s'il s'agit bien de pépites ou alors si ces règles sont liées à un bruit présent dans les données. Si ces règles sont dues au bruit alors l'introduction volontaire de bruit dans les données devrait les altérer et elles ne devraient plus être détectées par l'algorithme de recherche des pépites.

La connaissance du support minimum pour la résistance à un α-bruit permet de déterminer le support en dessous duquel ces règles instables apparaissent.

Nous obtenons ainsi un nouvel algorithme qui est très semblable à la méthode classique permettant de rendre les réseaux de neurones résistants au bruit (voir, par exemple, (Haykin 1998)).
Nommons $S_{\alpha_{noise}}$ le « support minimal pour la résistance à un α-bruit ».

Pour l'instant, l'implantation actuelle ne calcule le support minimum que pour les règles ayant exactement un seul attribut en prémisse.
Cette nouvelle heuristique améliore efficacement l'algorithme, pour une des bases de données étudiées, comme nous allons le voir dans la section suivante.

5.1 Validation expérimentale de l'approche

Nous avons testé l'algorithme 4 sur des données artificielles. Notre objectif est de montrer que l'impact des données bruitées sur les règles extraites est non négligeable et qu'il est difficile de proposer des solutions fiables pour éviter ces problèmes. La solution proposée pour réduire l'impact du bruit a été testée en suivant le protocole expérimental que nous présentons ci-après.

Dans les expérimentations réalisées, nous avons engendré des bases de données articielles en utilisant l'outil *IBMDataGen* disponible sur la page personnelle de M. Zaki [2]. Cet outil permet d'engendrer des bases de données transactionnelles. Le nombre de transactions, ainsi que le nombre d'attributs de la base sont paramétrables.

2. http://www.cs.rpi.edu/~zaki/software/

Algorithme 4 $DetecteS_{\alpha_{noise}}$

Entrée: $\mathcal{D}_n^p$: la base de données étudiée
Entrée: α_{noise} : bruit introduit dans les données
Entrée: N : nombre d'itérations pour obtenir le support minimal $S_{\alpha-noise}$
Sortie: $S_{\alpha_{noise}}$: support minimal garantissant la stabilité des règles les moins-contradictoires en présence d'un bruit α_{noise}

Début
$S_{\alpha_{noise}} = 0;\ \mathcal{E}_{noise} = \emptyset;\ S_{noise}^{prec} = 0;\ N = 1$
tant que $(S_{\alpha_{noise}} \neq S_{noise}^{prec})$ and $(N \leq 100)$ **faire**
$S_{noise}^{prec} = S_{\alpha_{noise}}$
si $(\mathcal{E}_{noise} = \emptyset)$ **alors**
$\mathcal{E}'_1 = ExtrairePepites(\mathcal{D}, 1, S_{\alpha_{noise}})$
$\mathcal{E}_{noise} = \mathcal{E}'_1$
sinon
$\mathcal{E}'_1 = ExtrairePepites(\mathcal{D}', 1, S_{\alpha_{noise}})$
$\mathcal{E}_{noise} = \mathcal{E}_{noise} \cap \mathcal{E}'_1$ -- *règles stables*
$\mathcal{E}^-_{noise} = \mathcal{E}_{noise} - \mathcal{E}'_1$ -- *règles disparues à cause du bruit*
$\mathcal{E}^+_{noise} = \mathcal{E}'_1 - \mathcal{E}_{noise}$ -- *règles apparues à cause du bruit*
$S^-_{max} = support(R)\ tq\ R \in \mathcal{E}^-_{noise}\ \wedge\ \forall R' \in \mathcal{E}^-_{noise}\ ,\ support(R) \geq support(R')$
$S_{\alpha_{noise}} = max(S_{\alpha_{noise}}, S^-_{max})$
fin si
si $(S_{\alpha_{noise}} \neq S_{noise}^{prec})$ **alors**
$N = 1$ -- *Supports différents : redémarrage d'une nouvelle boucle de 100 itérations*
sinon
$N = N+1$ -- *Supports identiques donc poursuite de la boucle courante de bruitage*
fin si
$\mathcal{D}' \Rightarrow IntroduireBruit - b(\mathcal{D}, \alpha_{noise})$
fin tant que
Retourner $S_{\alpha_{noise}}$
Fin

Pour chaque des bases de données engendrées, l'algorithme 1 d'extraction des pépites de connaissance a été appliqué avec les paramètres $K_{max} = 1$ et $min_{sup} = 0$. Nous avons testé différentes mesures de qualité en remplacant simplement, dans le cœur de l'algorithme, la mesure contramin par l'une des mesures suivantes: la Confiance, l'Intensité d'Implication Classique (Gras 1979), l'Intensité d'Implication Entropique (Gras et al. 2001), la Nouveauté (Lavrač et al. 1999) et l'Indice de Lœvinger (Lœvinger 1947). Nous nous sommes limités à ces mesures car nous pensons qu'elles illustrent suffisament d'aspects différents de la qualité des règles.

Les expérimentations ont été réalisées sur des bases de données contenant relativement peu d'attributs (en l'occurence 40 attributs booléens) et 30000 transactions.

Dix bases de données différentes ont été engendrées. Et pour chaque base de données obtenues, nous avons réalisé dix fois les expérimentations liées au bruit. Pour chaque expérimentation, le bruit introduit dans les données varie de 1% à 10%.

Le Tableau 5 présente les résultats obtenus pour ces différentes expérimentations sans la détection du support minimal $S_{\alpha_{noise}}$ et le Tableau 6 présente les résultats obtenus avec la détection du support minimal $S_{\alpha_{noise}}$.

Les Tableaux 5 et 6 s'interprètent de la façon suivante: pour une mesure de qualité, l'impact du bruit (avec ou sans détection du support) introduit dans les données (1%, 2%, 5% ou 10%) se manifeste de deux manières: les règles qui disparaissent **(-)** et les règles qui apparaissent **(+)**, par rapport aux règles obtenues à partir des données non bruitées.

Mesure	bruit observé en fonction du bruit introduit							
	1%		2%		5%		10%	
	- (%)	+(%)	-(%)	+ (%)	-(%)	+ (%)	- (%)	+(%)
Contramin	5,32	2,5	8,2	2,76	10,77	5,32	16,68	7,39
Confiance	15,98	6,52	22,96	8,55	31,57	14,2	42,55	22,9
Nouveauté	9,48	4,92	13,05	6,99	21,75	11,77	31,88	18,65
Indice de Lœvinger	12,96	8,3	19,3	10,47	27,95	16,83	40,33	29,76
IIC	10,22	17,64	13,43	24,19	19,73	35,54	26,1	45,07
IIE	11,51	12,19	13,85	17,23	19,2	25,15	30,35	33,52

TAB. 5 – *Résultats obtenus sur une base de données contenant* 30000 *transactions et* 40 *attributs, sans détection du support* $S_{\alpha_{noise}}$.

Mesure	bruit observé en fonction du bruit introduit							
	1%		2%		5%		10%	
	- (%)	+ (%)	- (%)	+ (%)	- (%)	+ (%)	- (%)	+ (%)
Confiance	9,94	13,72	21,25	32,57	35,63	46,47	44,81	68,7
Nouveauté	32,08	21,67	39,33	23	71,48	40,89	72,35	18,56
Indice de Lœvinger					11,54	35,38	28,0	50,99
IIC			12,38	26,19	19,3	35,44	24,19	45,82
IIE	15,42	19,58	12,49	27,52	17,02	32,9	28,46	54,33

TAB. 6 – *Résultats obtenus sur une base de données contenant* 30000 *transactions et* 40 *attributs, avec détection du support* $S_{\alpha_{noise}}$.

Pour la moindre contradiction, nous n'avons aucun résultat concernant l'impact de la détection du support $S_{\alpha_{noise}}$ car pour les diverses expérimentations effectuées, ce support était nul et n'engendrait donc aucun élagage. Nous ne pouvons donc rien conclure sur l'utilisation de $S_{\alpha_{noise}}$ pour la moindre contradiction.

Par contre, pour les autres mesures, nous pouvons voir que globalement le bruit observé pour les règles intéressantes qui disparaissent à diminuer de manière significative. L'approche retenue pour réduire l'impact du bruit provoque une augmentation dramatique du bruit lié au nombre de règles erronées qui sont détectées dans les données bruitées.

La solution que nous avons proposé permet donc bien de réduire le pourcentage de règles intéressantes et qui disparaissent lorsque les données sont bruitées, par contre, le pourcentage de règles erronées induites par cette approche rend celle-ci inutilisable car les règles purement dues au bruit sont difficles à détecter et seul l'expert peut les éliminer. Or le temps de l'expert est précieux et il est donc préférable de ne pas le solliciter avec des règles probablement erronées.

Nous proposons donc dans la section suivante une solution permettant d'évaluer la « fiabilité » d'une mesure lorsque le problème de l'extraction de pépites de connaissance est abordée en présence de données bruitées.

5.2 Évaluation de la « fiabilité » des mesures de qualité

N'oublions pas qu'un de nos objectifs majeurs est de minimiser le travail de l'utilisateur. Il est donc important de lui proposer un ensemble de règles qui vérifie au mieux les critères de qualité imposés par l'utilisateur et qui soit le plus résistant possible au bruit.

Nous proposons donc une approche visant, non pas à réduire l'impact du bruit dans les règles extraites, mais permettant d'associer à chaque règle proposée à l'expert une estimation de sa résistance au bruit.

Considérons une base de données $\mathcal{B}$ à partir de laquelle un ensemble de règles $\mathcal{E}$ est extrait (en utilisant la mesure de qualité m). Nous proposons d'introduire du bruit dans $\mathcal{B}$, puis à partir de la nouvelle base bruitée $\mathcal{B}'$, d'extraire le nouvel ensemble de règles $\mathcal{E}'$. Cette première étape, en tout point identique à celle déjà présentée dans les sections précédentes, est suivie d'une phase de comparaison des règles de $\mathcal{E}$ avec celles de $\mathcal{E}'$. L'objectif de cette comparaison est de vérifier si les règles trouvées à partir de $\mathcal{B}$ sont présentes dans $\mathcal{E}'$. Si tel est le cas, alors ces règles sont considérées comme fiables.

La répétition de cette opération permet d'obtenir pour chaque règle de $\mathcal{E}$ un pourcentage de « fiabilité » (par rapport au bruit).

L'algorithme 5 correspond à cette nouvelle approche.

Nous avons validé cette nouvelle approche sur le même type de données artificielles que celles utilisées précédemment. Nous nous sommes focalisés sur les pépites de connaissance contenant exactement un attribut en prémisse et un en conclusion. Les résultats obtenus permettent d'établir un classement entre les différentes mesures de qualité étudiées. Le classement obtenu est relatif à la fiabilité moyenne observée sur $N = 20$ itérations, et pour 10 bases de données différentes contenant 30000 transactions décrite par 50 attributs. Pour obtenir ce classement, nous avons bruité les données en introduisant entre 1 et 10% de bruit de la forme **b** dans les données.

Algorithme 5 Calcul de la fiabilité des règles.

Entrée:
$\mathcal{B}_n^p$: la base de données étudiée
α_{noise} : bruit introduit dans les données
N : nombre d'itérations pour obtenir le support minimal $S_{\alpha-noise}$
Sortie:
$\mathcal{E}$: ensemble de règles
Début
$\mathcal{E}_{noise} = \emptyset$ $N = 1$
$\mathcal{E} = ExtrairePepites(\mathcal{B}, 1, 0)$
−− on associe un compteur aux règles pour déterminer leur fiabilité
Pour tout $(R \in \mathcal{E})$ **faire**
$R.cpt \leftarrow 0$
fin pour
pour $(i = 1; i \leq N;\ i++)$ **faire**
$\mathcal{B}' = IntroduireBruit - b(\mathcal{B}, \alpha_{noise})$
$\mathcal{E}' = ExtrairePepites(\mathcal{B}', 1, 0)$
$\mathcal{E}_{stable} = \mathcal{E} \cap \mathcal{E}'$ *−− règles stables*
−− on incrémente le compteur associé aux règles
Pour tout $(R \in \mathcal{E}_{stable})$ **faire**
$R.cpt \leftarrow R.cpt + 1$
fin pour
fin pour
−− on normalise le compteur associé aux règles
Pour tout $(R \in \mathcal{E})$ **faire**
$R.cpt \leftarrow \frac{R.cpt}{N}$
fin pour
Retourner $\mathcal{E}$
Fin

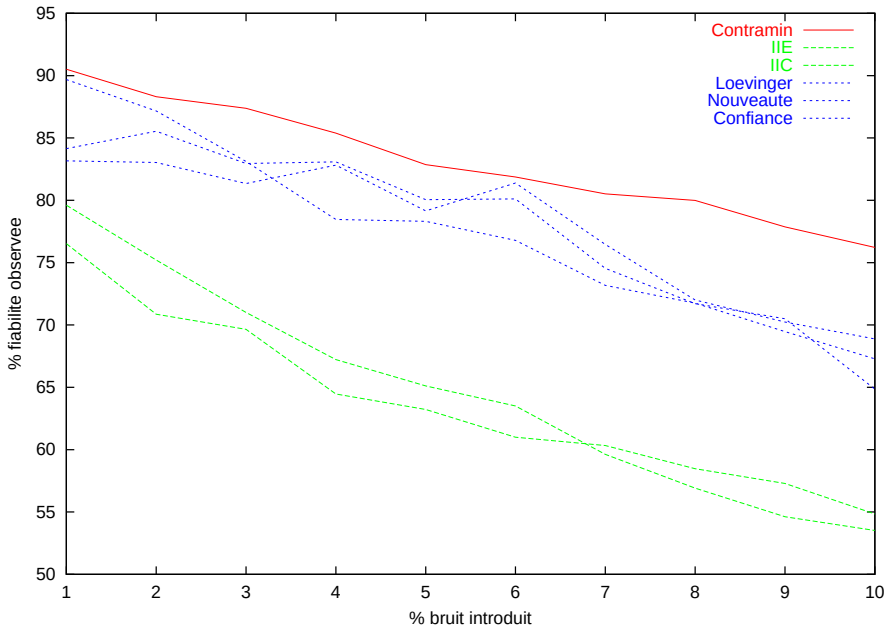

FIG. 3 – *Évolution de la fiabilité en fonction du bruit introduit.*

La figure 3 présente l'évolution de la fiabilité, pour les six mesures étudiées, en fonction du bruit introduit dans les données.

l'Intensite d'Implication Classique et Entropique sont les deux mesures les moins fiables lorsque les données sont bruitées. La Moindre Contradiction est la mesure la plus fiable. Et les autres mesures (Nouveauté, Lœvinger et Confiance) ont un comportement relativement proche lorsque les données sont bruités. Le classement de ces quatre mesures varie en fonction du bruit introduit dans les données.

Pour compléter ces résultats et valider l'approche retenue, nous avons aussi mesuré le pourcentage de règles qui, pour chaque mesure, sont présentes dans $\mathcal{E}$ et ne sont jamais retrouvées dans $\mathcal{E}'$ pour l'ensemble des 20 expérimentations réalisées. Seule la Moindre Contradiction arrive à retrouver sur l'ensemble des 20 expérimentations au moins une fois chaque règle de $\mathcal{E}$.

6 Conclusions et perspectives

L'extraction de « pépites » de connaissance dans des données corrélées est difficile voire impossible avec les approches classiques basées sur l'utilisation du support et de la confiance. L'approche proposée ici permet d'extraire de telles pépites, représentées par des règles d'association ayant de faibles supports et représentant les règles les moins contredites par les données.

L'extraction de ces règles est fondée sur l'utilisation du contexte local de l'en-

semble des règles en cours d'évaluation. Cette approche, classique dans le domaine de la fouille de données, permet ici de réduire significativement le nombre de règles proposées à l'expert. Son travail est donc allégé et les connaissances obtenues sont valides et intéressantes.

Les résultats obtenus ne représentent qu'une faible partie des connaissances cachées dans les données mais ces connaissances, souvent ignorées par les méthodes classiques, peuvent être utiles pour l'expert.

Les heuristiques d'élagage utilisées dans notre algorithme peuvent être appliquées aux règles obtenues à partir d'une approche classique. Cependant, le nombre de règles à analyser est très volumineux et, contrairement à notre approche, de nombreuses règles sont inutilement engendrées car finalement élaguées par les heuristiques utilisées.

La validation de notre approche doit être poursuivie sur des corpus de tailles plus importantes. De tels travaux sont en cours, en collaboration avec un expert, sur un corpus des liens-DNA des levures.

Concernant l'étude du bruit, les résultats obtenus sont assez encourageants, mais ils montrent que l'extraction de règles insensibles au bruit est très difficile à réaliser.

Nous avons montré, à travers les diverses approches envisagées et les expérimentations associées, que la prise en considération du bruit pouvant exister dans les données ou pouvant les altérer représente un aspect important de la qualité des connaissances extraites à partir des données.

Bien que nous ayons réduit notre étude à une forme de bruit relativement simpliste, nous avons constaté qu'il est difficile de proposer des solutions permettant de réduire l'impact du bruit sur les connaissances obtenues.

L'approche que nous avons proposée, fondée sur la détection automatique du support minimal permettant de ne plus perdre les meilleures règles lorsque les données sont bruitées, a permis de réduire significativement le pourcentage de règles disparaissant lorsque les données sont bruitées. Malheureusement, cette première approche a entraîné une augmentation dramatique du pourcentage de règle apparaissant lorsque les données sont bruitées et ce faisant augmente significativement le travail de l'expert pour trier les bonnes règles des règles erronées.

Nous avons donc opté pour une nouvelle approche dont l'objectif majeur est d'assister l'expert dans le choix et l'analyse des règles que nous lui proposons plutôt que de concevoir un filtre visant à sélectionner automatiquement les meilleures règles. Cette nouvelle approche permet d'associer à chaque règle détectée par l'algorithme d'extraction des pépites de connaissance dans les données un pourcentage de fiabilité correspondant à la résistance de la règle lorsque les données contiennent $\alpha_{noise}\%$ de bruit réparti de manière aléatoire. Les résultats obtenus sur des données aléatoires ont montré que, parmi les diverses mesures étudiées, la Moindre Contradiction est celle qui se comporte le mieux lorsque les données sont bruitées. En effet, la Moindre Contradiction est la mesure ayant le pourcentage moyen de résistance au bruit le plus élevé et ce pour toutes les valeurs de bruit introduit, $\alpha_{noise} \in [1..10]$.

La méthode proposée pour extraire les pépites de connaissance a été validée sur des données réelles et l'ajout de l'indicateur de fiabilité aux pépites de connaissance permet aux experts de mieux évaluer la qualité des résultats obtenus. Cette démarche doit encore être validée sur d'autres bases de données mais les premiers résultats sont

prometteurs.

Références

Agrawal, R., Imielinski, T., et Swami, A. N. (1993). Mining association rules between sets of items in large databases. Dans *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216.

Agrawal, R. et Srikant, R. (1994). Fast algorithms for mining association rules. Dans Bocca, J. B., Jarke, M., et Zaniolo, C., editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann.

Agrawal, R. et Srikant, R. (1995). Mining sequential patterns. Dans Yu, P. S. et Chen, A. S. P., editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan. IEEE Computer Society Press.

Azé, J. (2003). Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *RSTI série RIA-ECA*, 17(1-2-3):171–182.

Azé, J. et Kodratoff, Y. (2002a). évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. *Extraction des connaissances et apprentissage*, 1(4):143–154.

Azé, J. et Kodratoff, Y. (2002b). A study of the effect of noisy data in rule extraction systems. Dans *Proceedings of the Sixteenth European Meeting on Cybernetics and Systems Research (EMCSR'02)*, volume 2, pages 781–786.

Azé, J. et Roche, M. (2003). Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. *RSTI série RIA-ECA*, 17(1-2-3):283–294.

Blake, C. et Merz, C. (1998). UCI repository of machine learning databases.

Brin, S., Motwani, R., Ullman, J. D., et Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. Dans Peckham, J., editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 255–264. ACM Press.

Daudé, F. (1992). *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*. PhD thesis, Université de Rennes 1.

Freitas, A. A. (1998). On objective measures of rule surprisingness. Dans *Principles of Data Mining and Knowledge Discovery*, pages 1–9.

Gras, R. (1979). Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques. Master's thesis, Université de Rennes 1.

Gras, R., Kuntz, P., et Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille des données. *Revue Mathématique et Sciences Humaines*, 154-155:9–29.

Haykin, S. (1998). *Neural Networks - A Comprehensive Foundation*. Prentice Hall, 2nd edition.

Kodratoff Y., Azé J., Roche M. et Matte-Tailliez O. (2003). Des textes aux associations qu'ils contiennent. *numéro spécial RNTI "Entreposage et Fouille de données", JDS 2003*, 1:171-182

Kodratoff, Y. (2000). Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts.

Lavrač, N., Flach, P., et Zupan, B. (1999). Rule evaluation measures: A unifying view. Dans Džeroski, S. et Flach, P., editors, *Ninth International Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 of *Lecture Notes in Artificial Intelligence*, pages 174–185. Springer-Verlag.

Lœvinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61:1–49.

Sahar, S. (1999). Interestingness via what is not interesting. Dans *Knowledge Discovery and Data Mining*, pages 332–336.

Weiss, G. M. et Hirsh, H. (1998). The problem with noise and small disjuncts. Dans *Proc. 15th International Conf. on Machine Learning*, pages 574–578. Morgan Kaufmann, San Francisco, CA.

Summary

Most of the classical approaches for the extraction of association rules are based on the use of thresholds, set by the expert, to prune the search space. The choice of these thresholds, supposed to efficiently separate the set of interesting rules from the set of obvious rules, is quite difficult even for a domain expert. Considering that the data may be noisy and that the extraction of "nuggets" of knowledge (i.e., association rules with small support) may be of particular interest to the expert, then the classical methods are often unable to deal with this problem.

We propose a new association rule extraction measure called "least-contradiction". We show that this measure (i) enables us to extract "nuggets" of knowledge from the data, without drowning in a huge amount of rules having small supports, (ii) reacts somewhat less badly to noise than the other classical measures.

Keywords : Association rules, measures of quality, noise.