

Méthodes et mesures d'intérêt pour l'extraction de règles d'exception

Béatrice DUVAL*, Ansaf SALLEB**◇, Christel VRAIN**

*LERIA UFR Sciences 2 Bd Lavoisier 49045 Angers Cedex 01

Beatrice.Duval@univ-angers.fr,

<http://www.info.univ-angers.fr/pub/bd/>

**LIFO Rue Léonard de Vinci BP 6759 45067 Orléans Cedex 02

Ansaf.Salleb,Christel.Vrain@lifo.univ-orleans.fr

<http://www.univ-orleans.fr/SCIENCES/LIFO>

◇BRGM 3 Avenue Claude Guillemin BP 6009 Orléans Cedex 02

Résumé. Les systèmes de génération de règles sont en général fondés sur des critères leur permettant de juger de la qualité des règles engendrées. On recherche souvent les règles solides, avec un support et une confiance suffisants, *i.e.*, concernant une partie importante de la population et vérifiées sur un grand nombre d'individus. Cependant, les experts sont souvent plutôt intéressés par des règles qui les surprennent, soit parce qu'elles sont peu fréquentes (de support faible) mais de confiance élevée, soit parce qu'elles représentent des exceptions aux règles solides. Une recherche exhaustive des règles d'exceptions n'est pas envisageable, les stratégies de recherche sont alors tout aussi importantes que les critères de qualité. C'est pourquoi dans ce papier, nous présentons un état de l'art, non seulement des mesures, mais aussi des méthodes pour l'extraction d'exceptions.

1 Introduction

L'extraction de connaissances dans les bases de données est devenue aujourd'hui un domaine de recherche prometteur et très actif, permettant de découvrir des connaissances dans un ensemble de données volumineux. Dans ce papier, nous nous intéressons principalement aux connaissances exprimées sous forme de règles. Ces règles doivent être non seulement intelligibles à l'utilisateur, mais également intéressantes, exploitables, c'est-à-dire qu'elles doivent aider l'expert dans son travail et notamment dans la prise de décision. Beaucoup de travaux ont été effectués sur l'extraction des règles d'association, comme introduites dans [Agrawal *et al.*, 1993], et exprimant des corrélations dans les données.

Par la suite, on considère une base de données à une seule table sur des attributs qualitatifs, éventuellement binaires. On appelle *item* l'instantiation d'une valeur à un attribut et *itemset* un ensemble d'items. Un tuple de la base de données est un itemset qu'on appelle aussi *transaction*.

On note $\mathcal{I} = \{x_1, x_2, \dots, x_n\}$, l'ensemble des items possibles et $\mathcal{T} = \{t_1, t_2, \dots, t_m\}$ un ensemble de transactions.

Extraire des règles d'association consiste à rechercher toutes les règles *solides* de la forme $\mathcal{X} \rightarrow \mathcal{Y}$, où \mathcal{X} et \mathcal{Y} sont des itemsets disjoints ; une telle règle exprime le

fait que les items de \mathcal{Y} tendent à apparaître avec ceux de \mathcal{X} . La notion de solidité d'une règle n'est pas définie formellement, elle est mesurée par des critères, comme par exemple le couple $(s\%, c\%)$ donnant respectivement le support et la confiance de la règle. Le support de la règle $s\%$ représente la proportion de transactions de \mathcal{T} contenant \mathcal{X} et \mathcal{Y} . La confiance $c\%$ est la proportion de transactions parmi celles couvrant \mathcal{X} qui couvrent aussi \mathcal{Y} . Beaucoup de méthodes, comme l'algorithme initial *Apriori* [Agrawal *et al.*, 1993], se décomposent en deux étapes. La première consiste à extraire l'ensemble \mathcal{F} de tous les itemsets fréquents, par rapport à un support minimum *Min-Supp* fixé *a priori* par l'utilisateur. Dans un deuxième temps, l'extraction de règles se fait de façon effective sur \mathcal{F} : on considère chaque règle pouvant être engendrée par un itemset fréquent de \mathcal{F} ($\mathcal{X} \rightarrow \mathcal{Y}$ est engendrée si $\mathcal{X} \cup \mathcal{Y} \in \mathcal{F}$, $\mathcal{X} \cap \mathcal{Y} = \emptyset$) et on mesure sa qualité par un indice statistique, comme par exemple la confiance. Il en découle qu'une règle n'est générée que si l'itemset constitué des items de sa prémisse et de sa conclusion a un support suffisant. De ce fait, une règle de bonne qualité mais de faible support ne pourra jamais être détectée par cette approche. Ces approches ont le défaut d'engendrer un grand nombre de règles, ce qui rend leur exploitation par des experts difficile. De plus, beaucoup de ces règles avec un support élevé s'avèrent peu intéressantes, car déjà connues par l'expert, alors qu'il recherche souvent une connaissance surprenante, *exceptionnelle* dans les données. Ainsi, il est devenu fondamental de proposer de nouvelles méthodes permettant de n'engendrer que des règles susceptibles d'intéresser l'expert.

Dans cette optique, plusieurs travaux se sont intéressés à l'extraction de *règles d'exception*, citons [Piatetsky-Shapiro et Matheus, 1994, Silberschatz et Tuzhilin, 1995], [Liu *et al.*, 1999, Padmanabhan et Tuzhilin, 1998, Suzuki, 1999, Savasere *et al.*, 1998]. Une règle générale ou règle de sens commun (pouvant être une règle d'association) décrit une régularité assez forte, observée sur un grand nombre d'individus. Une règle d'exception représente, pour un nombre assez faible d'individus, une régularité qui contredit la règle de sens commun. Comme cette régularité concerne une population beaucoup plus faible, elle est en général moins connue et c'est cette notion de surprise qui fait l'intérêt des règles d'exception. Une telle règle apporte une connaissance plus fine des phénomènes observés dans les données.

Dans le domaine du crédit bancaire, par exemple, une règle de sens commun indique qu'une personne au chômage ne peut pas obtenir de crédit. Néanmoins, une exception stipule que les chômeurs ayant un projet solide de création d'entreprise peuvent obtenir un crédit dans ce but. Cet exemple illustre le schéma général d'exceptions qui nous intéressent. On se trouve en présence de deux règles, qui ont des conclusions contradictoires et qui peuvent toutes les deux être appliquées dans une même situation, celle d'un chômeur demandeur de crédit ayant un projet de création d'entreprise. L'une des règles, la règle d'exception, s'applique dans des conditions plus spécifiques (ses prémisses contiennent les prémisses de la règle de sens commun) et c'est cette règle que l'on souhaite appliquer plutôt que la règle de sens commun lorsque ces conditions spécifiques sont réunies.

La notion d'exceptions repose sur des itemsets non fréquents, et la recherche des exceptions ne peut donc être menée de manière exhaustive sur de grands volumes de données. C'est pourquoi elle se fait toujours en référence à des règles de sens commun.

On distingue deux écoles pour l'extraction des règles d'exception : l'approche *subjective* et l'approche *objective*. Elles se distinguent principalement par le fait que les règles de sens commun peuvent être soit données par l'utilisateur, soit extraites automatiquement, et par les critères utilisés pour mesurer l'importance des règles d'exceptions apprises.

Dans l'approche subjective, les connaissances du domaine données par l'expert constituent un système de règles de sens commun appelées convictions ¹; l'intérêt d'une règle d'exception est jugé en utilisant des critères subjectifs dépendant de l'utilisateur comme la *surprise* ² et l'*opérationabilité* ³ d'une règle qui caractérise les actions que l'on peut envisager quand une exception est trouvée.

L'approche objective n'utilise pas de connaissances *a priori* et mène de front la recherche des règles de sens commun et des exceptions associées. Les règles sont évaluées à l'aide de critères statistiques comme le support, la confiance, l'entropie, ... qui dépendent seulement des itemsets et des données qui les ont générées. Ces critères doivent permettre de juger à la fois l'intérêt de la règle de sens commun et de son exception. La difficulté de l'approche objective est qu'une exploration exhaustive de toutes les règles est très coûteuse, voire impossible. C'est pourquoi la plupart des travaux dans cette approche proposent une recherche partielle, en limitant la complexité syntaxique des règles étudiées. C'est le critère de réduction le plus simple à mettre en œuvre, nécessitant peu l'intervention de l'expert et justifié par l'argument qu'une règle contenant trop d'attributs est difficilement compréhensible.

Dans la mesure où l'on ne peut plus se baser sur le support pour élaguer l'espace des règles, des stratégies de recherche doivent être définies pour générer les règles et les évaluer. C'est la raison pour laquelle nous présentons dans ce papier à la fois les algorithmes qui génèrent les règles et les mesures de qualité utilisées.

Dans cet article, nous présentons différents travaux qui ont été menés sur le thème de l'extraction de règles d'association avec exception, en distinguant ceux relevant de l'approche objective (section 2) de ceux relevant de l'approche subjective (section 3). Dans la section 4, nous décrivons des travaux relevant de l'une et l'autre de ces deux approches. Enfin, la section 5 présente un travail consacré aux associations négatives; il se distingue des autres travaux, mais la problématique nous a semblé suffisamment proche et intéressante pour être abordée dans ce papier.

2 Approche Objective

2.1 Recherche non dirigée des couples règles-exceptions

Cette section présente les travaux développés par Suzuki [Suzuki et Shimura, 1996, Suzuki, 1996, Suzuki, 1997, Suzuki, 1999, Suzuki et Kodratoff, 1998]. La recherche non dirigée des exceptions consiste à rechercher simultanément une règle appelée règle de sens commun et une règle appelée règle d'exception décrivant des exceptions à cette

¹*belief* en anglais

²*unexpectedness* en anglais

³*actionability* en anglais

règle de sens commun. Dans la suite, une telle connaissance sera appelée *couple règle-exception*.

La recherche de couples règle-exception consiste à trouver des couples de règles de la forme

$$Y_\mu \rightarrow x$$

$$Y_\mu \wedge Z_\nu \rightarrow x'$$

où x et x' sont des items de même attribut mais portant sur des valeurs d'attributs différentes et $Y_\mu = y_1 \wedge y_2 \dots \wedge y_\mu$, $Z_\nu = z_1 \wedge z_2 \dots \wedge z_\nu$ sont des conjonctions d'items. Par exemple, x et x' peuvent être des items instanciant un attribut de classe par deux valeurs différentes, on a alors deux règles qui s'appliquent aux mêmes situations mais concluent sur des classes différentes.

Un des aspects centraux est de proposer un critère d'évaluation des différents couples règles-exception que l'on peut obtenir.

Le critère d'évaluation peut être donné par un unique indice qui évalue globalement la pertinence du couple de règles. Dans ce cas, le pré-ordre total sur les couples règle-exception induit par cet indice permet de fixer le nombre maximal de couples qui doivent être retournés par le processus de recherche.

Le critère d'évaluation peut également être composé de plusieurs indices, comme par exemple le support et la confiance, et pour chaque indice un seuil minimum est spécifié par l'utilisateur. Dans ce cas, on ne dispose plus de pré-ordre total sur les couples et il est plus difficile d'obtenir un nombre intéressant de couples règle-exception. En effet, des seuils inappropriés pour le jeu de données peuvent conduire à obtenir trop de couples ou au contraire aucun. Un processus de mise à jour des seuils doit donc être proposé.

Quel que soit le critère d'évaluation retenu, la recherche non dirigée des exceptions est vue comme l'exploration d'un espace de recherche où chaque nœud décrit un couple règle-exception. L'exploration de cet espace est nécessairement limitée.

La section 2.1.1 décrit brièvement l'algorithme de découverte non dirigée des exceptions. La section 2.1.2 décrit les différents critères d'évaluation qui ont été proposés dans les travaux de Suzuki.

2.1.1 Algorithme

Un couple règle-exception

$$Y_\mu \rightarrow x$$

$$Y_\mu \wedge Z_\nu \rightarrow x'$$

est considéré comme un nœud $r(\mu, \nu)$ dans un arbre de recherche. Les nœuds de profondeur 1 de l'arbre sont les couples de conclusions (x, x') que l'on doit considérer. Ces nœuds peuvent donc être vus comme des règles sans prémisses donc pour lesquelles $\mu = 0$ et $\nu = 0$. Lorsque la profondeur augmente de 1, un item est ajouté soit dans les prémisses de la règle de sens commun, soit dans les prémisses de la règle d'exception. Un nœud de profondeur 2 vérifie $\mu = 1$ et $\nu = 0$ et correspond à une règle de sens commun avec une seule prémisse. Un nœud de profondeur 3 vérifie donc $\mu = 1$ et $\nu = 1$ et plus généralement un nœud de profondeur l ($l \geq 4$) vérifie donc $\mu + \nu = l - 1$

($\mu, \nu \geq 1$). Par conséquent, un nœud $r(\mu', \nu')$ descendant du nœud $r(\mu, \nu)$ correspond à une règle pour laquelle $\mu' \geq \mu$ et $\nu' \geq \nu$. Cet arbre est exploré par une recherche en profondeur d'abord, bornée par une profondeur maximum M . Par conséquent on ne considère que les couples de règles pour lesquels $\mu + \nu \leq M - 1$; pour fixer les idées, les expériences présentées dans les articles se réfèrent à une profondeur maximum de 8. Comme on l'a déjà dit, une telle contrainte syntaxique est justifiée par le fait que les experts appréhendent mal des règles contenant beaucoup de prémisses. Pour chaque critère d'évaluation étudié, Suzuki propose des propriétés qui permettent un élagage lors de l'exploration de l'arbre.

2.1.2 Critères d'évaluation des couples règles-exception

Les premiers systèmes MEPRO et MEPROUX proposés par Suzuki [Suzuki, 1996, Suzuki et Shimura, 1996] utilisent un indice issu de la théorie de l'information pour déterminer l'intérêt d'une règle. Pour caractériser l'intérêt d'une règle induite à partir d'un ensemble de données, Smyth et Goodman [Smyth et Goodman, 1992] ont défini la *J-measure* aussi appelée *entropie compressée moyenne* (Average Compressed Entropy, notée ACE dans la suite). Etant donné Y_μ et x tels que $p(x|Y_\mu) \geq 0.5$, l'ACE de la règle $Y_\mu \rightarrow x$ est définie par :

$$ACE(x, Y_\mu) = p(x, Y_\mu) \log_2 \left(\frac{p(x|Y_\mu)}{p(x)} \right) + p(\neg x, Y_\mu) \log_2 \left(\frac{p(\neg x|Y_\mu)}{p(\neg x)} \right)$$

que l'on peut aussi écrire :

$$ACE(x, Y_\mu) = p(Y_\mu) * [p(x|Y_\mu) \log_2 \left(\frac{p(x|Y_\mu)}{p(x)} \right) + p(\neg x|Y_\mu) \log_2 \left(\frac{p(\neg x|Y_\mu)}{p(\neg x)} \right)]$$

L'ACE est donc composée du facteur $p(Y_\mu)$ qui évalue la généralité de la règle et du terme $p(x|Y_\mu) \log_2 \left(\frac{p(x|Y_\mu)}{p(x)} \right) + p(\neg x|Y_\mu) \log_2 \left(\frac{p(\neg x|Y_\mu)}{p(\neg x)} \right)$ que l'on notera ici $j(x, Y_\mu)$. Le terme $j(x, Y_\mu)$ apparaît dans la littérature sous le nom d'entropie croisée ou mesure de Kullback-Lieber, qui permet d'évaluer la dissimilarité entre deux distributions de probabilités. Ici le terme $j(x, Y_\mu)$ évalue la différence entre notre connaissance *a priori* sur x et notre connaissance *a posteriori* connaissant Y_μ .

Une règle d'exception $Y_\mu \wedge Z_\nu \rightarrow x'$ dont l'ACE est forte peut ne pas être intéressante si l'ACE de la règle de sens commun associée est faible. C'est pourquoi l'indice retenu pour évaluer l'intérêt d'un couple règle-exception doit tenir compte des ACE de chaque règle. Suzuki [Suzuki, 1996] montre que la moyenne arithmétique ne convient pas pour évaluer l'intérêt combiné des deux règles; en effet, pour $p(x)$ et $p(x')$ fixés, la moyenne arithmétique des ACE atteint son maximum lorsque $ACE(x, Y_\mu) = 0$ ou lorsque $ACE(x', Y_\mu \wedge Z_\nu) \approx 0$, ce qui ne correspond pas à des cas intéressants que l'on veut retenir. La moyenne géométrique des ACE, noté GACE, ne présente pas cet inconvénient et elle est donc retenue comme indice pertinent de l'intérêt du couple règle-exception :

$$GACE(x, Y_\mu, x', Z_\nu) = \sqrt{ACE(x, Y_\mu) \cdot ACE(x', Y_\mu \wedge Z_\nu)}$$

De plus, une règle $Y_\mu \wedge Z_\nu \rightarrow x'$ ne peut pas être considérée comme une exception surprenante à la règle de sens commun $Y_\mu \rightarrow x$ si la règle $Z_\nu \rightarrow x'$, qui est appelée *règle de référence*, possède une forte confiance, c'est-à-dire si $p(x'|Z_\nu)$ est élevée. Pour tenir compte de cet élément, Suzuki indique avoir exploré plusieurs contraintes possibles et sans justifier ce choix, il précise que dans le système MEPROUX, $p(x'|Z_\nu)$ doit satisfaire la contrainte suivante :

$$p(x'|Z_\nu) \leq p(x') + \frac{1 - p(x')}{2}$$

Nous verrons dans la section 2.2 une autre manière de tenir compte de la règle de référence .

De plus, pour que la combinaison de Y_μ et Z_ν soit réellement significative, on impose que

$$p(x'|Y_\mu, Z_\nu) > p(x'|Z_\nu)$$

Cela permet de trouver des couples règle-exception qui ont à la fois un caractère intéressant et pour lesquels la règle d'exception est effectivement inattendue.

On donne dans le tableau 1 un exemple de règles extraites du jeu de données "mush-room" [Murphy et Aha, 1995], où la classe comestible (e) ou vénéneux (p) est le seul attribut autorisé en conclusion (d'autres expériences sans restriction sur les items de conclusion ont également été menées). Ce tableau donne dans l'ordre la règle de sens commun, la règle d'exception dans laquelle les prémisses de la règle de sens commun sont notées \mathcal{H} et la règle de référence.

	$p(x Y)$	$p(x)$	$p(Y)$	ACE	GACE
bruis=f, ring-number=o \rightarrow class=p	0.74	0.48	0.54	0.107	
\mathcal{H} , ss-aring=f \rightarrow class=e	1.00	0.52	0.05	0.048	0.0713
ss-aring=f \rightarrow class=e	0.74	0.52	0.07		

TAB. 1 – Un exemple de couple règle-exception et la règle de référence associée

D'après ce couple règle-exception, 74% des champignons tels que "bruis=f, ring-number=o" sont vénéneux, mais 100% d'entre-eux sont en fait comestibles si "ss-aring=f". Cette exception ne peut pas être prédite par la règle de référence qui a une probabilité conditionnelle de seulement 74%.

Dans ses autres travaux [Suzuki, 1997, Suzuki, 1999], Suzuki utilise plusieurs critères pour juger de l'intérêt d'un couple règle-exception. On considère toujours un couple règle de référence et règle d'exception de la forme :

$$Y_\mu \rightarrow x$$

$$Y_\mu \wedge Z_\nu \rightarrow x'$$

Pour définir la qualité des règles, l'utilisateur fournit 5 seuils θ_1^S , θ_1^F , θ_2^S , θ_2^F et θ_2^I et les règles retenues doivent satisfaire les conditions suivantes :

$$p(Y_\mu) \geq \theta_1^S \tag{1}$$

$$p(x|Y_\mu) \geq \theta_1^F \tag{2}$$

$$p(Y_\mu, Z_\nu) \geq \theta_2^S \quad (3)$$

$$p(x'|Y_\mu, Z_\nu) \geq \theta_2^F \quad (4)$$

$$p(x'|Z_\nu) \leq \theta_2^I \quad (5)$$

Les seuils θ_1^S (qui concerne le support) et θ_1^F (qui concerne la probabilité conditionnelle) garantissent la généralité et la confiance de la règle de sens commun. De même, les seuils θ_2^S et θ_2^F garantissent la généralité et la confiance de la règle d'exception. La règle d'exception s'observe sur peu de cas mais possède une très bonne confiance, c'est pourquoi on s'attend à ce que l'utilisateur donne des seuils vérifiant $\theta_1^S > \theta_2^S$ et $\theta_1^F < \theta_2^F$. L'équation (5) garantit que la règle de référence $Z_\nu \rightarrow x'$ n'est pas une régularité forte, ce qui est nécessaire pour que les exceptions soient pertinentes.

Les expériences menées montrent que cette méthode conduit à des règles fiables mais qui ne sont pas toujours surprenantes. Dans [Suzuki et Kodratoff, 1998], l'intensité d'implication est utilisée pour trouver des règles d'exception plus surprenantes. L'intensité d'implication [Gras et Lahrer, 1993] est un critère qui permet de mesurer le caractère inattendu d'une règle. On mesure en fait le degré de surprise lié au fait que la règle possède si peu de contre-exemples.

Pour appliquer l'intensité d'implication dans le cadre d'un couple règle-exception, on va s'intéresser au fait que la règle d'exception a peu de contre-exemples dans un univers restreint aux éléments qui satisfont la prémisse de la règle de sens commun. Parmi les couples de règles qui satisfont les équations (1-5), on retient donc les couples pour lesquels l'intensité d'implication de la règle d'exception est la plus forte.

Les papiers que nous avons étudiés n'indiquent pas si des comparaisons ont été faites entre l'évaluation par un seul indice entropique ou l'évaluation à l'aide de seuils. Mais il faut bien sûr noter que la méthode utilisant plusieurs seuils pose le problème du choix de seuils convenables ; cela demande à la fois une expertise sur les données et sur le processus de découverte, et nécessite donc une implication forte de l'utilisateur. Des seuils trop exigeants vont empêcher la découverte de règles d'exception, et au contraire, on risque d'obtenir beaucoup trop de couples règles-exceptions si on relâche trop les valeurs de ces seuils. Pour résoudre cette difficulté, [Suzuki, 1999] propose une méthode de mise à jour dynamique des seuils afin de garantir que l'on trouve un nombre fixé η de couples de règles.

2.2 Prise en compte de la règle de référence dans la mesure d'intérêt

Le travail présenté dans [Hussain *et al.*, 2000] suit une approche similaire mais s'appuie surtout sur un algorithme de recherche très différent.

Les auteurs travaillent dans le cadre d'attributs booléens et s'intéressent à des règles d'exception qui correspondent au schéma suivant :

$Y_\mu \rightarrow x$	règle de sens commun	support élevé, confiance élevée
$Y_\mu \wedge Z_\nu \rightarrow \neg x$	règle d'exception	support faible, confiance élevée
$Z_\nu \rightarrow \neg x$	règle de référence	support faible et/ou confiance faible

En fait, la règle de référence est remplacée par une règle de sens commun associée. En effet, si $Z_\nu \rightarrow x$ est une règle de sens commun, qui possède donc un support et une

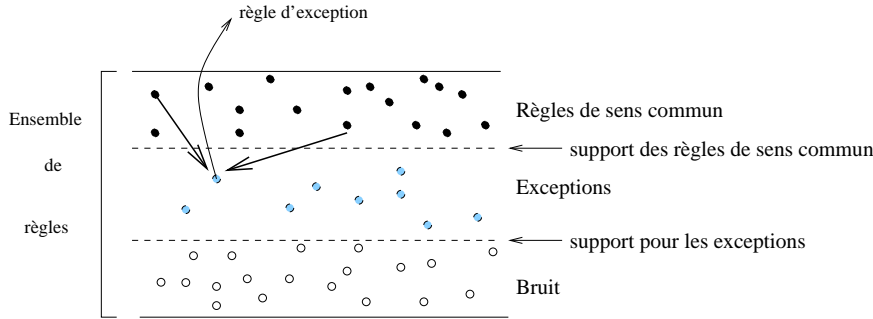


FIG. 1 – ensemble de règles

confiance qui dépassent les seuils requis, alors la règle $Z_\nu \rightarrow \neg x$ est considérée comme une règle de référence.

L'algorithme proposé recherche les exceptions en s'appuyant sur les règles de sens commun. L'utilisateur spécifie les seuils minimum pour le support et la confiance des règles de sens commun. Par un algorithme comme Apriori, on calcule d'abord tous les itemsets fréquents et toutes les règles de sens commun. Pour chaque couple de règles de sens commun de la forme $(Y_\mu \rightarrow x, Z_\nu \rightarrow x)$, si $Y_\mu \cup Z_\nu$ n'est pas un itemset fréquent, alors $Y_\mu, Z_\nu \rightarrow \neg x$ est une règle candidate pour être une exception (voir figure 1). L'ensemble des règles candidates étant construit, on examine une nouvelle fois les données pour déterminer le support de chaque règle candidate et ne retenir que celles qui dépassent un seuil fixé par l'utilisateur pour les règles d'exception. Les règles extraites à partir des mesures de support et de confiance sont alors filtrées pour ne retenir que celles ayant une mesure d'intérêt suffisante. La mesure d'intérêt proposée cherche à évaluer la règle d'exception $Y_\mu, Z_\nu \rightarrow \neg x$ vis à vis des deux règles de sens commun $(Y_\mu \rightarrow x, Z_\nu \rightarrow x)$. Pour cela, les auteurs utilisent l'entropie croisée ou distance de Kullback Leibler (voir 2.1.2). Rappelons que l'entropie croisée entre deux lois de probabilités $p(x)$ et $q(x)$ est définie par

$$D(p(x)||q(x)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Cette mesure, non symétrique, s'interprète comme l'inadéquation à supposer que la distribution est $q(x)$ quand elle est en fait $p(x)$ [Smyth et Goodman, 1992]. Pour comparer la distribution associée à la règle $Y_\mu, Z_\nu \rightarrow \neg x$ avec les distributions associées aux deux règles $Y_\mu \rightarrow x, Z_\nu \rightarrow x$, les auteurs proposent donc d'ajouter les entropies croisées $D(Y_\mu, Z_\nu \rightarrow \neg x || Y_\mu \rightarrow x)$ et $D(Y_\mu, Z_\nu \rightarrow \neg x || Z_\nu \rightarrow x)$. De plus, ils veulent comparer les informations apportées par ces règles à la fois pour ce qui concerne les confiances et les supports. L'intérêt relatif aux confiances RI_c est défini en considérant dans la formule de l'entropie croisée les probabilités conditionnelles, ce qui donne donc :

$$RI_c = p(x|Y_\mu, Z_\nu) \log_2 \frac{p(x|Y_\mu, Z_\nu)^2}{p(x|Y_\mu)p(x|Z_\nu)} + p(\neg x|Y_\mu, Z_\nu) \log_2 \frac{p(\neg x|Y_\mu, Z_\nu)^2}{p(\neg x|Y_\mu)p(\neg x|Z_\nu)}$$

De même, la mesure d'intérêt liée aux support, RI_s , est la somme de deux entropies croisées calculées à partir des probabilités jointes, ce qui donne :

$$RI_s = p(x, Y_\mu, Z_\nu) \log_2 \frac{p(x, Y_\mu, Z_\nu)^2}{p(x, Y_\mu)p(x, Z_\nu)} + p(\neg x, Y_\mu, Z_\nu) \log_2 \frac{p(\neg x, Y_\mu, Z_\nu)^2}{p(\neg x, Y_\mu)p(\neg x, Z_\nu)}$$

L'intérêt total est alors $RI = RI_c + RI_s$.

Ce travail est à comparer à celui de [Suzuki, 1996]. Nous avons vu, en section 2.1.2, que [Suzuki, 1996] utilise un unique indice appelé $GACE(x, Y_\mu, x', Z_\nu)$ pour évaluer un couple règle-exception; cette mesure combine les ACE des règles $Y_\mu \rightarrow x$ et $Y_\mu \wedge Z_\nu \rightarrow x'$, mais on avait vu qu'une contrainte supplémentaire devait être imposée sur $p(x', Z_\nu)$. Pour améliorer cette approche, la mesure d'intérêt utilisée dans [Hussain *et al.*, 2000] intègre également la règle de référence $Z_\nu \rightarrow x'$ dans l'évaluation, ce qui évite d'avoir à déterminer la contrainte appropriée sur $p(x', Z_\nu)$. On constate d'ailleurs, par expérimentation sur des jeux de données, que l'indice $GACE$ et la mesure d'intérêt se comportent de la même manière si l'on applique la mesure d'intérêt sans avoir de règle de référence; plus précisément ces deux mesures classent les règles extraites dans le même ordre. Lorsqu'on dispose des règles de référence, la mesure d'intérêt permet une évaluation plus fine des règles en proposant dans un unique indice une évaluation du triplet (règle de sens commun, exception, règle de référence). Ainsi l'article présente des exemples de règles obtenues à partir de jeux de données du site d'Irvine. On constate sur ces exemples que deux couples règle-exception, ayant les mêmes valeurs pour l'indice $GACE$, peuvent être départagés par l'indice RI qui prend en compte leurs règles de référence. Cet indice cherche à rendre compte à la fois de l'intérêt et du caractère inattendu des exceptions.

3 Approche Subjective

D'après Piatetsky-Shapiro et Matheus dans [Piatetsky-Shapiro et Matheus, 1994], les mesures objectives, malgré leur fondement statistique, n'arrivent pas de façon générale à exprimer l'intérêt des schémas obtenus par un processus d'extraction de connaissances (un schéma désigne ici une connaissance extraite, soit un itemset, soit une règle). Des mesures subjectives sont donc nécessaires. Ces mesures ne dépendent pas seulement de la forme des schémas et des données qui les ont engendrés, mais dépendent aussi de l'utilisateur qui va examiner ces schémas. C'est une façon d'impliquer l'utilisateur, car un schéma s'il est intéressant pour un utilisateur peut ne pas l'être pour un autre. Silberschatz et Tuzhilin [Silberschatz et Tuzhilin, 1996] identifient deux critères permettant de juger si un schéma est intéressant ou non du point de vue subjectif :

- la surprise : un schéma est dit intéressant s'il surprend l'utilisateur, *i.e.* si ce schéma contredit ce qu'il croit être une évidence dans ses données,
- l'opérationabilité : un schéma est intéressant s'il aide l'utilisateur à prendre une "bonne" décision dans son travail.

Bien que l'opérationabilité soit un critère majeur, il apparaît que cette notion est difficile à définir formellement. C'est pourquoi les auteurs proposent de capturer cette

notion à travers le critère de surprise, considérant que les schémas qui poussent le plus l'utilisateur à agir sont plutôt les schémas surprenants. La surprise d'un schéma est reliée à un système de convictions \mathcal{C} de l'utilisateur : ce schéma sera d'autant plus surprenant qu'il perturbe \mathcal{C} . Un système de convictions \mathcal{C} peut être modélisé à l'aide de formules logiques pouvant exprimer des convictions de deux types : *convictions fermes* et *convictions souples*. Les convictions fermes traduisent des connaissances qui sont toujours vraies et qui ne peuvent être changées par de nouvelles observations. Si on détecte alors des schémas contradictoires à ce genre de convictions, cela signifie que des erreurs existent dans les données. Par contre, les convictions souples sont celles que l'utilisateur peut modifier ou faire évoluer selon les circonstances. Une mesure d'importance, appelée *degré de conviction*, est associée à chaque conviction souple et traduit à quel point l'utilisateur croit en cette conviction. Les auteurs proposent plusieurs approches pour affecter un degré à une conviction ; elles diffèrent par leur généralité et les propriétés qu'elles vérifient. Parmi les approches existantes, nous ne détaillons ici que l'approche bayésienne, que les auteurs considèrent comme la plus appropriée dans ce cadre.

Soit $\alpha \in \mathcal{C}$ une conviction, et ξ une évidence constituée, par exemple, des données et des convictions fermes. Le degré de α relativement à ξ , $d(\alpha|\xi)$, est donné par la probabilité conditionnelle $P(\alpha|\xi)$ que α soit valide étant donnée l'évidence ξ . Soit une nouvelle évidence E , le degré de la conviction α est mis à jour en utilisant la formule de Bayes suivante :

$$d(\alpha|E, \xi) = P(\alpha|E, \xi) = \frac{P(E|\alpha, \xi)P(\alpha|\xi)}{P(E|\alpha, \xi)P(\alpha|\xi) + P(E|-\alpha, \xi)P(-\alpha|\xi)}$$

À la découverte d'un nouveau schéma, on se demande si ce dernier perturbe notre système de convictions et à quel point il l'affecte, autrement dit s'il s'agit d'une exception ou non. Pour quantifier cette notion, une fonction de poids w_i est associée à chaque conviction souple α_i de \mathcal{C} . Ces poids sont normalisés pour que $\sum_{\alpha_i \in \mathcal{C}} w_i = 1$. La formule suivante mesure l'intérêt d'un schéma p relativement à un système de convictions souples \mathcal{C} et une évidence ξ :

$$I(p, \mathcal{C}, \xi) = \sum_{\alpha_i \in \mathcal{C}} w_i |d(\alpha_i|p, \xi) - d(\alpha_i|\xi)|$$

D'après cette définition, les nouveaux schémas intéressants sont ceux qui modifient le plus les convictions que l'on avait. Cette approche semble bien adaptée aux situations où l'on gère des données qui changent dans le temps, ce qui est le cas des systèmes de transactions en ligne, tels que les systèmes de réservation de places d'avions, de transactions bancaires, de marketing . . .

Les auteurs proposent alors une méthode pour réaliser un système de découverte guidé par les croyances, qui s'appuie sur la propriété suivante : s'il existe dans \mathcal{C} une conviction α telle que $d(\alpha|E, \xi) \neq d(\alpha|\xi)$, où E représente un ensemble de nouvelles données, alors il existe dans E un schéma p tel que $I(p, \mathcal{C}, \xi) \neq 0$. Lorsque de nouvelles données sont disponibles, les degrés de croyance sont revus pour prendre en compte ces nouvelles informations ; si l'on constate alors que les degrés de croyance ont beaucoup évolué (au delà d'un seuil fixé), cela signifie que les nouvelles données contiennent des motifs intéressants et il est alors pertinent de déclencher un processus d'extraction. Notons que la notion de surprise n'est pas utilisée ici seulement pour l'extraction de règles avec

exceptions, mais aussi pour détecter l'évolution d'un système de convictions. Nous n'avons pas trouvé d'article relatant la mise en œuvre de cette idée.

La principale difficulté dans cette proposition d'approche subjective réside dans le calcul des degrés de convictions. En effet, le calcul des probabilités *a posteriori* à partir des probabilités *a priori* en utilisant la formule de Bayes nécessite de nombreux calculs qui peuvent rendre cette approche trop lourde à mettre en œuvre. Néanmoins nous trouvons très intéressante l'idée de piloter le processus de découverte de connaissances en observant l'évolution des degrés de croyances. Dans les systèmes où beaucoup de données sont produites chaque jour, il peut être très utile d'offrir de tels indicateurs qui, en signalant une évolution notable dans les données, vont inciter les décideurs à lancer une phase d'extraction de connaissances.

Les idées développées dans [Silberschatz et Tuzhilin, 1996] sont intéressantes, mais sont difficiles à appliquer, car elles placent l'utilisateur au cœur du processus.

Utiliser des convictions a été repris par dans [Padmanabhan et Tuzhilin, 1998] et dans [Padmanabhan et Tuzhilin, 2000]. Ils ont proposé des algorithmes de découverte de règles utilisant le support et la confiance comme principales mesures pour rechercher dans les données des règles *inattendues* ou *exceptionnelles* qui contredisent les convictions de l'expert. Dans [Padmanabhan et Tuzhilin, 1998], une règle d'exception est définie comme suit :

Soit $X \rightarrow Y$ une règle de conviction. La règle $A \rightarrow B$ est une règle d'exception par rapport à la conviction $X \rightarrow Y$ sur la base de données \mathcal{D} , si les conditions suivantes sont vérifiées :

1. B et Y se contredisent mutuellement,
2. $X \wedge A$ est fréquent dans \mathcal{D} ,
3. $A, X \rightarrow B$ est une règle solide. Puisque B et Y se contredisent, il s'en suit que $A, X \rightarrow \neg Y$ est une règle solide.

L'algorithme *ZoomUR*⁴ proposé dans [Padmanabhan et Tuzhilin, 2000] est composé de deux parties :

- *ZoominUR* est un algorithme qui explore les règles d'exception à la manière de l'algorithme Apriori [Agrawal *et al.*, 1993] :

Étant donné une base de données et un système de règles de convictions de la forme $X \rightarrow Y$, *ZoominUR* découvre toutes les règles d'exceptions de la forme $X, A \rightarrow C$, telles que C contredit Y et telles que leurs supports et confiances sont suffisants par rapport à des seuils fixés *a priori*. Pour chaque conviction $X \rightarrow Y$, *ZoominUR* génère d'abord de manière incrémentale tous les itemsets fréquents pouvant potentiellement générer des exceptions. À chaque itération de *ZoominUR*, les itemsets sont générés comme suit : à la k -ième itération, sont générés les itemsets de la forme $\{X, P, C\}$ où C contredit Y . Notons que pour déterminer la confiance de la règle $X, P \rightarrow C$, les supports de $\{X, P, C\}$ et $\{X, P\}$ doivent être déterminés. Une fois que tous les itemsets fréquents sont retrouvés, *ZoominUR* génère les règles inattendues par rapport à la conviction considérée. Ces règles contiennent toutes X dans leurs prémisses.

⁴pour Zoom to Unexpected Rules

- *ZoomoutUR* recherche des règles plus générales aux règles d'exception trouvées, i.e., de la forme $X' \wedge A \rightarrow \neg Y$ avec $X' \subset X$.

L'algorithme *ZoomUR* a été testé sur deux bases de données réelles, dont l'une est issue du domaine commercial et concerne les habitudes d'achats de clients. Par exemple, étant donnée la conviction suivante : "les gens qui travaillent font leurs courses le week end", le processus a permis de trouver les deux exceptions :

- "En décembre, les gens qui travaillent tendent à faire leurs courses dans la semaine".
- "Les gens qui travaillent et qui ont des enfants tendent à faire leurs courses dans la semaine".

Ces deux exceptions sont intéressantes ; la première peut être expliquée par le nombre important de courses à effectuer pendant la période des fêtes et aux vacances scolaires et la seconde pourrait être justifiée par le fait que les courses doivent être faites souvent dans une famille nombreuse.

Pour cet exemple, la deuxième phase de l'exploration, à l'aide de l'algorithme *ZoomoutUR* a extrait la règle suivante :

- "En décembre, les acheteurs tendent à faire leurs courses dans la semaine".

On voit ici que cette règle est bien une généralisation de la première exception trouvée. En revanche, il ne s'agit plus d'une exception à la conviction donnée initialement. De plus, on peut dire que si, en décembre tous les acheteurs tendent à faire leurs courses en semaine, alors la première exception trouvée (en décembre, les gens qui travaillent tendent à faire leurs courses dans la semaine) n'est plus réellement intéressante puisqu'elle est une spécialisation d'une règle plus générale.

Nous retrouvons ici une idée déjà présentée dans les sections précédentes ; en effet on a vu que les travaux qui s'appuient sur des mesures objectives utilisent non seulement une règle de sens commun mais aussi une règle de référence pour ne conserver que des exceptions intéressantes. Ici l'idée est appliquée différemment puisque, partant d'une conviction initiale, des exceptions, finalement peu intéressantes, sont un élément intermédiaire pour obtenir des connaissances plus générales en rapport avec la conviction initiale (courses le week-end ou courses dans la semaine, par exemple) mais qui ne la contredisent pas nécessairement.

La base de données commerciale utilisée dans ces tests expérimentaux contient près de 90 000 tuples définis sur 36 attributs discrets et où chaque attribut possède entre 2 et 12 modalités différentes. Pour 15 règles de convictions considérées, *ZoomUR* a généré plus de 600 règles d'exception contre 40 000 règles générées avec l'algorithme Apriori pour les mêmes seuils de support et confiance.

Bien que ces tests montrent que le nombre de règles générées par *ZoomUR* est nettement inférieur au nombre de règles générées par l'algorithme Apriori, beaucoup de règles parmi les règles d'exception trouvées peuvent être redondantes car les règles de convictions sont considérées indépendamment les unes des autres. C'est pourquoi les auteurs ont par la suite étendu leur algorithme *ZoominUR* à l'algorithme *MinZoomUR* proposé dans [Padmanabhan et Tuzhilin, 2000], pour l'extraction d'ensembles minimaux de règles d'exception.

On voit bien l'intérêt de cette approche pour limiter le nombre de règles engendrées, en se focalisant sur les règles intéressantes pour l'expert, et liées à ses convictions. En

revanche, la principale critique que l'on peut faire sur ce travail est la pauvreté des mesures de qualité utilisées. De plus, l'algorithme tel qu'il est présenté ne semble pas optimal, car il nécessite le calcul de nombreux supports.

4 Approche Mixte

Nous appelons approche mixte une approche où les connaissances de sens commun peuvent être soit fournies par l'utilisateur, soit calculées par une méthode inductive. Ce contexte se retrouve dans d'autres situations étudiées en intelligence artificielle. C'est le cas notamment dans les travaux qui cherchent à appliquer des méthodes inductives pour maintenir une base de règles, tout en préservant une bonne compréhension des connaissances par un expert humain [Saux *et al.*, 2002].

Dans [Liu *et al.*, 1999], Liu et al. proposent une approche simple d'extraction de règles d'exceptions solides basée sur l'analyse des déviations. Les connaissances de sens commun pour lesquelles on cherche des exceptions sont données ou sont calculées par une autre approche d'apprentissage. L'approche proposée se décompose en 4 étapes :

- **Étape 1** : Identification des attributs intéressants en se basant sur des connaissances du domaine ou sur un choix de l'utilisateur ou encore, en appliquant une méthode d'induction telle que la recherche des associations. Cette phase de filtrage permet de focaliser sur quelques attributs seulement. Si on considère le domaine du crédit bancaire, par exemple, on peut s'intéresser aux deux variables booléennes *emploi* indiquant si la personne possède ou non un emploi et *crédit* indiquant si un crédit a été accordé ou non.
- **Étape 2** : Construction de tables de contingences pour chaque couple d'attributs jugé intéressant. Si on considère le couple d'attributs (A, B) , on supposera que la variable A peut prendre les valeurs A_1, A_2, \dots, A_r et la variable B les valeurs B_1, B_2, \dots, B_s , la table de contingence associée est donnée par le tableau 2 ci-dessous :

A	B				total
	B_1	B_2	\dots	B_s	
A_1	$(n_{11})x_{11}$	$(n_{12})x_{12}$	\dots	$(n_{1s})x_{1s}$	$n_{1.}$
A_2	$(n_{21})x_{21}$	$(n_{22})x_{22}$	\dots	$(n_{2s})x_{2s}$	$n_{2.}$
\dots	\dots	\dots	\dots	\dots	\dots
A_r	$(n_{r1})x_{r1}$	$(n_{r2})x_{r2}$	\dots	$(n_{rs})x_{rs}$	$n_{r.}$
total	$n_{.1}$	$n_{.2}$	\dots	$n_{.s}$	n

TAB. 2 – Table de contingence des attributs A et B

Dans la table de contingence, les x_{ij} sont les fréquences d'une occurrence, *i.e.* d'un couple de valeurs (A_i, B_j) dans les données.

n est le total des fréquences, $n = \sum_{i=1}^r \sum_{j=1}^s x_{ij}$

$n_{i.}$ représente la fréquence marginale horizontale, $n_{i.} = \sum_{j=1}^s x_{ij}$

$n_{.j}$ représente la fréquence marginale verticale, $n_{.j} = \sum_{i=1}^r x_{ij}$

$n_{ij} = \frac{n_{i.} * n_{.j}}{n}$ est la fréquence attendue pour l'occurrence (A_i, B_j) .

La déviation par rapport à la valeur attendue est donnée par $\delta_{ij} = \frac{x_{ij} - n_{ij}}{n_{ij}}$

Ainsi pour l'exemple cité plus haut, la table de contingence associée au couple de variables (crédit,emploi) sera donnée par le tableau 3.

Crédit	Emploi		total
	oui	non	
non	(35.5) 28	(4.5) 12	40
oui	(75.5) 83	(9.5) 02	85
total	111	14	125

TAB. 3 – Table de contingence des attributs crédit et emploi

- **Étape 3** : Calcul des déviations δ_{ij} et identification des déviations négatives remarquables. Les fréquences attendues sont celles que l'on observerait en cas d'indépendance des deux attributs représentés dans la table de contingence. On se fixe un seuil de déviation pour ne retenir que les écarts suffisamment significatifs. Parmi les déviations significatives, les déviations positives correspondent à des itemsets fréquents qui confirment le lien attendu entre les attributs étudiés. En revanche, les déviations négatives correspondent à des itemsets peu fréquents qui peuvent révéler des exceptions. Dans l'exemple du crédit bancaire, le calcul des déviations est donné dans le tableau 4.

Emploi	Credit	x_{ij}	n_{ij}	δ
non	non	12	4.5	+1.68
non	oui	02	9.5	-0.79
oui	non	28	35.5	-0.21
oui	oui	83	75.5	+0.10

TAB. 4 – Déviations pour l'exemple Emploi,crédit

Pour un seuil de déviation $\delta_t = -0.5$, la seule déviation négative retenue correspond à l'occurrence (emploi=non, crédit=oui). Cet ensemble non fréquent correspond à un lien inattendu entre l'absence d'emploi et l'attribution de crédit, qui mérite que l'on vérifie si une exception est associée à cette occurrence. Cette vérification est l'objet de l'étape suivante.

- **Étape 4** : Extraction des exceptions. Pour cela, on ne considère que les données vérifiant l'occurrence de la déviation négative retenue dans l'étape précédente; ces données forment une "fenêtre" sur laquelle on peut appliquer une méthode comme *APriori* pour chercher les itemsets fréquents. Dans la fenêtre concernant les attributs $(A = A_i, B = B_j)$, tous les itemsets fréquents ainsi trouvés contiennent nécessairement $A = A_i$ et $B = B_j$ et conduisent à des règles de la forme, $(A = A_i) \wedge X \rightarrow (B = B_j)$ de confiance 1. Pour obtenir des règles d'exception fiables, on ne garde que les règles de cette forme où X est un itemset n'ayant pas un support trop grand dans le jeu de données complet.

Pour l'exemple du crédit, une règle d'exception trouvée est que "les femmes sans emploi ayant peu d'expérience se voient accorder un crédit". Cette règle a une confiance de 70% mais est peu fréquente dans la totalité des données.

Cette méthode diffère des précédentes dans la mesure où elle est guidée par une table de contingence. D'autres travaux [Brin *et al.*, 1997] ont montré l'intérêt des tables de contingence et du test du Chi-Deux pour la recherche de corrélations entre attributs. Néanmoins, contrairement aux arguments avancés par Brin et al. dans [Brin *et al.*, 1997], la méthode se focalise ici sur l'étude individuelle des cellules de la table. De plus, il est curieux de constater que cette méthode n'utilise pas de support minimum pour les exceptions étudiées; en effet tous les couples $(A = A_i, B = B_j)$ auxquels on s'intéresse sont ceux qui dévient sensiblement des fréquences attendues en cas d'indépendance des attributs. On risque donc d'examiner des couples dont la fréquence est très faible. Il serait par conséquent intéressant de voir comment se comporte cette méthode face à des données bruitées.

5 Associations négatives

Dans tous les travaux que nous venons de voir, on recherche des règles d'exception à des règles de sens commun. Une règle d'exception de la forme $X \rightarrow Y$ concerne un itemset $X \cup Y$ peu fréquent, mais possède une forte confiance, ce qui ne peut se produire que si l'itemset X est lui-même peu fréquent. Le travail présenté dans [Savasere *et al.*, 1998] s'intéresse lui aussi à des itemsets $X \cup Y$ non fréquents mais qui sont formés d'itemsets X et Y tous deux fréquents. Comme les combinaisons d'attributs qui ont une fréquence faible est potentiellement très grand, la recherche des combinaisons rares intéressantes se fait en étudiant les déviations par rapport aux itemsets fréquents, compte tenu d'une taxonomie d'un domaine. En fait, une exception consiste ici à observer une absence d'association entre des items, alors que des items qui leur sont proches d'après la taxonomie sont fortement associés. Les deux points qui nous semblent donc intéressants sont l'utilisation de connaissances sur le domaine sous forme de taxonomies et une recherche des itemsets non fréquents guidée par les itemsets fréquents et les connaissances.

Les auteurs s'intéressent donc aux règles d'associations dites *négatives* qui spécifient les items qu'un client a peu de chances d'acheter, sachant qu'il achète un certain ensemble d'items. Par exemple, il peut être intéressant de savoir que 60 % des clients qui achètent de l'eau gazeuse n'achètent pas de yaourt nature.

Rechercher de telles règles revient à rechercher les ensembles d'items qui ont peu de chance d'être achetés ensemble. Dans ce cadre, beaucoup de règles peuvent être engendrées dont un grand nombre risque d'être inintéressantes. Un des problèmes qui se pose est donc la détermination de mesures d'intérêt d'une règle négative. On peut mesurer le caractère inattendu d'une règle : une règle est intéressante si elle contredit ou si elle dévie de manière significative de nos attentes fondées sur nos croyances antérieures. Cependant, en l'absence de connaissances sur les préférences d'achat des consommateurs et si on suppose que tous les items sont achetés indépendamment les uns des autres, alors le support attendu d'une paire d'items risque d'être très faible. Dans ce cas, même si l'on se restreint aux ensembles d'items dont le support réel est

nul, la déviation relativement au support attendu sera très faible. Notons que cette hypothèse d'indépendance n'est en général pas vérifiée.

Pour résoudre ces problèmes et déterminer les règles dont on peut espérer un degré élevé d'associations positives et dont le support réel est beaucoup plus petit, les auteurs proposent d'utiliser des connaissances sur le domaine, sous forme de taxonomies qui permettent de regrouper les items similaires. L'idée sous-jacente est que des items qui ont un même parent dans la taxonomie devraient avoir le même type d'association avec les autres items. Par exemple, si des produits similaires A et B sont commercialisés sous deux marques différentes, et si A a une association très forte avec un autre item C , on pourrait s'attendre à ce que B ait le même type d'association avec C . Si ce n'est pas vérifié, on a une association négative intéressante entre B et C .

Considérons par exemple les connaissances suivantes [Savasere *et al.*, 1998] :

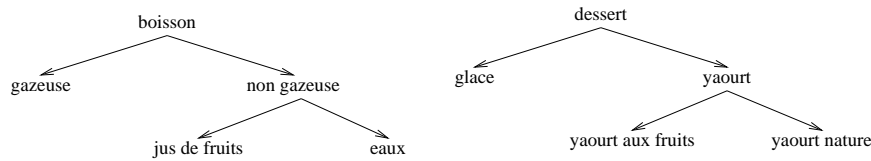


FIG. 2 – Une taxonomie sur des articles

Si les consommateurs qui achètent des boissons non gazeuses tendent à acheter des yaourts, on peut penser qu'il en sera de même des consommateurs qui achètent des jus de fruits. Une règle d'association négative intéressante sera une règle qui contredira cet apriori. On s'intéresse donc aux règles dont on peut attendre un degré élevé d'associations positives et dont le support réel est beaucoup plus petit.

Plus formellement, une *règle d'association négative* est une implication de la forme $X \not\rightarrow Y$ où X et Y sont des ensembles d'items tels que $X \cap Y = \emptyset$. La *mesure d'intérêt d'une règle d'association négative* $X \not\rightarrow Y$ est définie par :

$$RI = \frac{\mathcal{E}[\text{support}(X \cup Y)] - \text{support}(X \cup Y)}{\text{support}(X)}$$

où $\mathcal{E}[\text{support}(X \cup Y)]$ est le support attendu pour $X \cup Y$.

RI est " négativement relié " au support réel de $X \cup Y$: il est plus élevé si le support réel vaut 0 et il vaut 0 si le support réel est égal au support attendu.

L'algorithme proposé repose sur les points suivants :

- On fait l'hypothèse que les items qui ont un même parent dans la taxonomie ont des associations similaires avec les autres items.
- On recherche des règles d'associations négatives telles que le support de X et le support de Y sont supérieurs à $MinSup$ et l'intérêt de la règle est supérieur à $MinRI$.
- On se restreint aux règles négatives dont le support attendu peut être calculé à partir de la taxonomie. Par exemple, le support de $\{p_1, \dots, p_k\}$ où chaque p_i est un fils direct de \hat{p}_i est évalué comme suit :

$$\mathcal{E}[\text{sup}(p_1 \cup \dots \cup p_k)] = \frac{\text{sup}(\hat{p}_1 \cup \dots \cup \hat{p}_k) \times \text{sup}(p_1) \times \dots \times \text{sup}(p_k)}{\text{sup}(\hat{p}_1) \times \dots \times \text{sup}(\hat{p}_k)}$$

Considérons que pour l'exemple précédent, les supports sont donnés par le tableau 5.

<i>yaourt nature</i>	20000
<i>yaourt aux fruits</i>	10000
<i>yaourt</i>	30000
<i>eaux</i>	8000
<i>jus de fruits</i>	12000
<i>non gazeuse</i>	20000
<i>yaourt</i> \wedge <i>non gazeuse</i>	15000

TAB. 5 – Support des différents itemsets

Connaissant le support de l'itemset *yaourt* \wedge *non gazeuse*, on applique une règle de proportionnalité pour estimer le support d'un itemset formé d'un fils de *yaourt* et d'un fils de *non gazeuse*. Ainsi le support attendu de *yaourt nature* \wedge *eaux* est égal à :

$$15000 \times \frac{20000}{30000} \times \frac{8000}{20000} = 4000,$$

Si le support réel de *yaourt nature* \wedge *eaux* a une valeur beaucoup plus faible, égale à 800 par exemple, alors la règle d'association négative *eaux* \nrightarrow *yaourt nature* a une mesure d'intérêt de $\frac{4000-800}{8000} = 0.4$ et elle est intéressante si le seuil d'intérêt minimal est fixé à 0.4.

La critique que l'on peut faire à ce travail est principalement que la notion de règle d'association négative qui semble vraiment originale par rapport aux autres travaux n'est pas bien spécifiée, car sa sémantique n'est définie que par la mesure d'intérêt qui ne précise pas comment une telle règle pourrait être utilisée. De plus, l'approche repose sur la notion de taxonomie ; de telles connaissances ne sont pas toujours disponibles, la construction de telles taxonomies dépend de l'expert et de l'application qui a motivé leur élaboration.

6 Conclusion

Les différents travaux présentés dans cette synthèse traitent de la recherche de règles qui traduisent une régularité forte entre les items mais qui concernent un nombre relativement restreint d'individus. Ces règles appelées règles d'exception sont définies relativement à une règle appelée règle de sens commun qu'elles contredisent. Le schéma général qui décrit les travaux présentés dans les sections 2, 3 et 4 est donc le suivant. On s'intéresse à des couples de règles de la forme :

$$\begin{array}{ll} X \rightarrow Y & \text{appelée règle de sens commun notée RSC} \\ X \wedge A \rightarrow Z & \text{appelée règle d'exception notée EXC} \end{array}$$

où X et A sont des ensembles d'items et Y et Z sont des items contradictoires. Dans le cadre booléen, des items contradictoires correspondent à un attribut et sa négation, et dans le cas d'attributs discrets, des items contradictoires sont obtenus à partir d'un même attribut auquel sont affectées des valeurs différentes.

De plus, certains travaux considèrent également une troisième règle

$A \rightarrow Z$ appelée règle de référence notée REF

Les caractéristiques des différentes méthodes que nous avons présentées ici peuvent être résumées dans le tableau 6, qui utilise les notations RSC, EXC et REF pour désigner les règles.

Pour l'utilisation pratique de ces méthodes, il importe de distinguer entre l'approche objective et l'approche subjective.

L'approche objective ne nécessite aucune connaissance du domaine car elle découvre à la fois les règles de sens commun et les règles d'exception. On peut soit mener simultanément la recherche d'une règle de sens commun et d'une exception associée, soit au contraire extraire d'abord toutes les règles de sens commun par une méthode de type Apriori, puis chercher les exceptions à ces règles. Cette approche objective ne s'appuie que sur les données ; par conséquent elle a besoin de mesures pertinentes pour qualifier les couples de règles qui sont extraits.

Mis à part les travaux [Suzuki, 1997, Suzuki, 1999] où plusieurs seuils de support et de confiance sont utilisés, les différentes mesures proposées sont finalement assez semblables, il s'agit de mesures de nature entropique cherchant à évaluer à la fois l'intérêt et la surprise d'un couple de règles. Ces mesures ont été appliquées sur de jeux de données classiques du site d'Irvine, mais les articles ne relatent pas d'expérience sur des jeux de données artificiels permettant de voir, par exemple, comment se comportent ces mesures face à des jeux de données bruitées. Cela pourrait être intéressant puisque les exceptions sont des règles de support faible. Dans l'approche subjective, les règles de sens commun correspondent à un ensemble de connaissances, appelées convictions, données par un expert du domaine. On demande également à l'expert de qualifier son degré de croyance dans chacune de ses convictions. Si des croyances sûres sont remises en cause par les données observées, cela conduit à douter de la fiabilité des données. Si des croyances "souples" sont contredites par les données, on a trouvé des règles d'exception intéressantes car elles constituent des règles nouvelles que l'expert ne maîtrisait pas et qui vont lui permettre d'affiner sa compréhension du domaine. Comme la recherche des exceptions se fait par rapport aux règles données par l'expert, on pourrait dire que les pépites ne seront découvertes que si elles se trouvent dans les roches que l'expert connaît bien.

En fait, l'idée sous-jacente à tous les travaux que nous avons présentés est que les associations fréquentes risquent d'être déjà connues de l'utilisateur. Par conséquent, la recherche de connaissances intéressantes s'oriente vers les associations ayant un support relativement faible. Comme cette recherche ne peut être menée exhaustivement par les méthodes de type Apriori, tous les travaux s'appuient sur des connaissances déjà existantes ou déjà calculées (les règles de sens commun) pour limiter leur exploration. Nous devons citer néanmoins que des travaux se démarquent de cette approche et essaient d'exhiber des règles qui peuvent avoir un support faible, sans référence à des règles déjà données. C'est ce qui est proposé dans [Azé, 2003] où une nouvelle mesure de qualité d'une règle permet de mener une recherche n'utilisant pas de seuil de support minimum.

Enfin on peut souligner que traiter le problème des exceptions aux règles de sens commun n'est qu'un aspect d'un problème plus général qui est celui de la cohérence

	Informations que doit fournir l'utilisateur	Méthode	Mesures
Suzuki & al.	Seuils de support et de confiance pour RSC et EXC <i>ou</i> Aucune information	Exploration systématique (mais limitée) des couples (RSC,EXC)	Support et confiance <i>ou</i> Un unique indice entropique pour le couple (RSC, EXC)
Hussain & al.	Seuils de support et de confiance pour RSC et REF Seuil de support pour EXC	Apriori pour calculer RSC et REF Génération des EXC candidates Evaluation des triplets obtenus (retour aux données)	Un unique indice entropique pour évaluer (RSC, EXC, REF)
Padmanabhan & al.	Conviction RSC Seuil de support Seuil de confiance	Calcul des fréquents pouvant être exception à la conviction Règle retenue si confiance > seuil Généralisation des EXC pour trouver des règles REF	Support Confiance
Liu & al.	Couples d'attributs intéressants donnés ou calculés Seuil de déviation	Recherche des déviations négatives fortes par rapport à l'indépendance Recherche de règles fortes parmi les tuples vérifiant cette déviation Pas de règles RSC, ni de règles REF	Pas de mesure particulière
Savasere & al.	Seuil de support Seuil d'intérêt Taxonomie	Calcul de RSC $A \rightarrow B$ Recherche des règles négatives $X \not\rightarrow Y$, avec X proche de A et Y proche de B dans la taxonomie	Mesure de déviation entre support attendu et support réel

TAB. 6 – Caractéristiques des méthodes étudiées, RSC désigne une règle de sens commun, EXC, une règle d'exception, et REF une règle de référence

globale d'un ensemble de règles. En effet, on pourrait également se poser la question de savoir si un ensemble de règles extraites par un algorithme classique de type Apriori contient des contradictions, c'est-à-dire des schémas tels que ceux que nous avons vus dans cette étude. Bien sûr, cela ne peut arriver que si l'on travaille avec des seuils de support assez bas (20 % par exemple). Dans ce cas, il serait tout à fait intéressant de souligner à l'utilisateur ces rapprochements entre règles contradictoires.

Un autre thème de recherche qui nous semble intéressant est celui de l'apprentissage de règles d'association avec négation de la forme $L_1, \dots, L_p \rightarrow L_{p+1}, \dots, L_n$, où chaque L_i est un littéral, *i.e.*, un atome ou la négation d'un atome. Peu de travaux s'y sont intéressés et le problème est souvent résolu en introduisant dans la liste des variables un atome et sa négation et en les traitant comme des variables différentes. Cette solution n'est pas satisfaisante d'une part, car la taille du problème en nombre de variables est multipliée par deux, ce qui augmente la complexité et d'autre part, car les attributs A et $\neg A$ sont traités comme des attributs indépendants.

Ce problème est abordé partiellement dans les travaux que nous venons de citer puisque les règles avec exceptions introduisent la négation dans le membre gauche de la règle. Les travaux sur la recherche de règles d'associations négatives $X \not\rightarrow Y$ sont aussi intéressants dans ce cadre, mais ils posent vraiment des questions de sémantique des règles générées et en particulier, on peut se demander quelles relations peuvent exister entre la règle $X \not\rightarrow Y$ et la règle $X \rightarrow \neg Y$.

Références

- [Agrawal *et al.*, 1993] Rakesh Agrawal, Tomasz Imielinski, et Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman et Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.
- [Azé, 2003] J. Azé. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *Extraction des connaissances et apprentissage*, 17(1) :171–182, 2003.
- [Brin *et al.*, 1997] Sergey Brin, Rajeev Motwani, et Craig Silverstein. Beyond market baskets : Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, volume 26,2 of *SIGMOD Record*, pages 265–276, New York, May13–15 1997. ACM Press.
- [Gras et Lahrer, 1993] R. Gras et A. Lahrer. L'implication statistique : une nouvelle méthode d'analyse des données. *Mathématiques, Informatique et Sciences Humaines*, 120 :5–31, 1993.
- [Hussain *et al.*, 2000] Hussain, Liu, Suzuki, et Lu. Exception rule mining with a relative interestingness measure. In *PAKDD : Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*. LNCS, 2000.
- [Liu *et al.*, 1999] Huan Liu, Hongjun Lu, Ling Feng, et Farhad Hussain. Efficient search of reliable exceptions. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 194–203, 1999.

- [Murphy et Aha, 1995] P. M. Murphy et D. W. Aha. *UCI Repository of Machine Learning Databases*. Machine-readable collection, Dept of Information and Computer Science, University of California, Irvine, 1995. [Available by anonymous ftp from ics.uci.edu in directory pub/machine-learning-databases].
- [Padmanabhan et Tuzhilin, 1998] Balaji Padmanabhan et Alexander Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Knowledge Discovery and Data Mining*, pages 94–100, 1998.
- [Padmanabhan et Tuzhilin, 2000] Balaji Padmanabhan et Alexander Tuzhilin. Small is beautiful : discovering the minimal set of unexpected patterns. In *Knowledge Discovery and Data Mining*, pages 54–63, 2000.
- [Piatetsky-Shapiro et Matheus, 1994] Gregory Piatetsky-Shapiro et Christopher J. Matheus. The interestingness of deviations. In *KDD-94*, pages 25 – 36, 1994.
- [Saux et al., 2002] Elisabeth Le Saux, Philippe Lenca, et Philippe Picouet. Dynamic adaptation of rules bases under cognitive constraints. *European Journal of Operational Research*, 136 :299–309, 2002.
- [Savasere et al., 1998] Ashoka Savasere, Edward Omiecinski, et Shamkant B. Navathe. Mining for strong negative associations in a large database of customer transactions. In *ICDE*, pages 494–502, 1998.
- [Silberschatz et Tuzhilin, 1995] Abraham Silberschatz et Alexander Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Knowledge Discovery and Data Mining*, pages 275–281, 1995.
- [Silberschatz et Tuzhilin, 1996] Abraham Silberschatz et Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *Ieee Trans. On Knowledge And Data Engineering*, 8 :970–974, 1996.
- [Smyth et Goodman, 1992] P. Smyth et R. M. Goodman. An information theoretic approach to rule induction from databases. *Ieee Trans. On Knowledge And Data Engineering*, 4 :301–316, 1992.
- [Suzuki et Kodratoff, 1998] E. Suzuki et Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. In Jan M. Żytkow et Mohamed Quafafou, editors, *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-98)*, volume 1510 of *LNAI*, pages 10–18, Berlin, September 23–26 1998. Springer.
- [Suzuki et Shimura, 1996] Einoshin Suzuki et Masamichi Shimura. Exceptional knowledge discovery in databases based on information theory. In Evangelos Simoudis, Jia Wei Han, et Usama Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 275–278. AAAI Press, 1996.
- [Suzuki, 1996] Einoshin Suzuki. Discovering unexpected exceptions : A stochastic approach. In *Proceedings of RFSD96*, pages 225–232, 1996.
- [Suzuki, 1997] Einoshin Suzuki. Autonomous discovery of reliable exception rules. In David Heckerman, Heikki Mannila, Daryl Pregibon, et Ramasamy Uthurusamy, editors, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, page 259. AAAI Press, 1997.

[Suzuki, 1999] Einoshin Suzuki. Scheduled discovery of exception rules. In Setsuo Arikawa et Koichi Furukawa, editors, *Proceedings of the 2nd International Conference on Discovery Science (DS-99)*, volume 1721 of *LNAI*, pages 184–195, Berlin, December 6–8 1999. Springer.

Summary

The systems that generate association rules are usually based on criteria that enable to measure the quality of the learned rules. In the classical framework, support and confidence are used to evaluate the rules : a strong rule concerns a large proportion of the population and is satisfied by many individuals. However, experts are often interested by surprising rules, either because they are infrequent but with a high confidence, or because they represent exceptions to strong rules. Searching such rules cannot be achieved by reducing the support, because it would generate too many rules. This explains why many works rely on common sense rules, either given by the user or learned and search interesting exceptions to these rules. In this paper, we make a survey of existing methods and measures proposed for generating exception rules. Let us notice that for mining exception rules, an exhaustive search is not possible, and therefore methods and measures are tightly linked.