Découverte d'associations quantitatives générales et atypiques

Sylvie Guillaume*, Leïla Nemmiche-Alachaher*, Michel Schneider*

* Laboratoire LIMOS, UMR 6158 CNRS, Université Blaise Pascal
Complexe scientifique des Cézeaux, 63177 AUBIERE Cedex - France
{sylvie.guillaume, nemmiche, michel.schneider}@isima.fr

Résumé. Dans ce papier, nous proposons un nouveau type d'association mettant en jeu au moins une variable quantitative. Il s'agit de rechercher dans une population d'individus, les catégories qui s'écartent significativement du comportement normal de cette population. Plus exactement nous recherchons les catégories d'individus qui sont sur-représentées ou au contraire sous-représentées pour les fortes ou les faibles valeurs de la variable cible. Pour caractériser l'association, nous avons utilisé et étendu une mesure existante, l'intensité d'inclination. Comme toutes les catégories d'individus (*ou associations de variables*) ne sont pas d'égal intérêt pour l'utilisateur, ce type de connaissance permet, dans un premier temps, d'avoir une vision globale de l'association. Dans un deuxième temps il est possible de rechercher les intervalles de valeurs pour lesquels les écarts interviennent. Cette recherche des intervalles s'appuie sur les tableaux de contingence des écarts entre la situation observée et la situation attendue.

1 Introduction

La découverte des règles d'association Agrawal et al. (1993), Mannila et al. (1994) et Agrawal et al. (1996)) repose généralement sur deux mesures de fréquence (*support et confiance*) qui sont peu adaptées dans le cas d'analyses de données quantitatives. Il n'est pas possible de segmenter directement les données et une étape préliminaire de discrétisation est nécessaire.

Srikant et Agrawal (1996) ont proposé une technique basée sur une discrétisation automatique des données en maîtrisant la perte d'information engendrée par cette étape de préparation. Cependant Bay (2001a) et Ludl et Widmer (2000) ont montré qu'une discrétisation selon uniquement la distribution de la variable, sans tenir compte du contexte, peut conduire à des solutions non optimales. Mehta et Parthasarathy (2005) ont donc proposé une discrétisation qui prend en compte la distribution de chacune des variables mises en jeu. Lee et al. (2004) proposent une méthode augmentant la confiance accordée dans les motifs fréquents obtenus après discrétisation des variables quantitatives. Quant à Kuok et al. (1998), Zhang (1999) et Subramanyam et Goswami (2006), ils utilisent la technique des ensembles flous et Miller et Yang (1997) et Tong et al. (2005) utilisent des clusters pour la recherche des règles d'association quantitatives. D'autres auteurs optimisent

les mesures : Fukuda et al. (1996) optimisent le support et la confiance alors que Brin et al. (2005) optimisent le gain (*différence entre le nombre d'individus vérifiant la règle et le nombre d'individus vérifiant la prémisse*). Rückert et al. (2004) recherchent un nouveau type de règles pour les variables quantitatives à partir d'hyperplans et obtiennent, non plus des règles composées d'une conjonction de variables catégorielles ou d'intervalles (*règles à partir d'hyperrectangles de variables discrètes*), mais des règles du type : si la somme pondérée d'un ensemble de variables quantitatives est supérieure à un seuil donné, alors une autre somme pondérée de variables sera plus grande qu'un autre seuil avec une confiance digne d'intérêt. Georgii et al. (2005) ont appliqué ce nouveau type de règles à des données biologiques et plus particulièrement à des données d'expression des gênes. Salleb-Aouissi et al. (2007) recherchent des règles d'association quantitatives en utilisant un algorithme génétique qui découvre les meilleurs intervalles des variables quantitatives en optimisant le support et la confiance d'une règle.

Auman et Lindell (1999) proposent deux nouveaux types de règles en s'appuyant sur la distribution des valeurs des variables quantitatives : le premier type se compose uniquement de variables catégorielles en prémisse et le second, d'une variable numérique en prémisse. La variable conclusion est une variable quantitative (*ou un ensemble de variables quantitatives dans le premier type de règles*). Un exemple de règle du premier type donné par Auman et Lindell (1999) est : sexe = féminin $\rightarrow$ moyenne(salaire) = 7,90 $ par heure (*sachant que la moyenne de la variable salaire dans la base d'apprentissage est de 9,02 $ par heure*). D'après Auman et Lindell (1999), une règle est jugée intéressante si la catégorie d'individus vérifiant la prémisse a une moyenne pour la variable quantitative conclusion significativement différente du reste de l'ensemble d'apprentissage. Cet écart devant être suffisamment important, une différence minimum doit être spécifiée par l'utilisateur.

Webb (2001) étend les règles proposées par Auman et Lindell (1999) et les nomme règles d'impact. En effet, la recherche du premier type de règle de Auman et Lindell (1999) reposant sur la recherche des motifs fréquents, est très coûteuse dans le cas de données denses. De plus, leur approche les libère de la contrainte arbitraire de support minimum et repose sur un ensemble de mesures dont la mesure d'impact qui prend en compte non seulement la déviation de la variable conclusion mais également la couverture de la règle.

Nous proposons un nouveau type d'association mettant en jeu une variable quantitative que nous nommerons association quantitative générale et atypique. Ce type d'association recherche les catégories d'individus qui sont sur-représentées ou au contraire sous-représentées pour les fortes ou faibles valeurs de la variable quantitative cible. Voici un exemple de ce type d'association : les individus exerçant une profession libérale sont sur-représentés parmi les individus touchant un fort salaire et au contraire, sous-représentés parmi les individus percevant un faible salaire. Ainsi, ce type d'association recherche les catégories d'individus qui s'écartent significativement du comportement général de la population (*d'où l'appellation d'associations atypiques*), et plus particulièrement pour les fortes et les faibles valeurs de la variable quantitative.

La recherche de ces groupes d'individus sur-représentés pour une situation donnée s'apparente à la recherche des associations positives et inversement la recherche des catégories sous-représentées est similaire à la recherche des associations négatives. Wu et al. (2004) et Antonie et Zaane (2004) présentent des méthodes pour extraire des règles d'association positives et négatives et Yuan et al. (2002) décrivent une stratégie pour extraire des règles négatives en utilisant la connaissance du domaine.

Dans notre approche, nous nous libérons de la contrainte arbitraire de support minimum ainsi que de la recherche des motifs fréquents et contrairement à Webb (2001), nous utilisons une mesure normalisée pour extraire ces associations. Cependant, nous ne recherchons pas, dans un premier temps, les intervalles des variables quantitatives où l'association est jugée intéressante. Nous préférons acquérir une connaissance générale de l'association en se limitant uniquement à la notion de fortes et faibles valeurs pour les variables quantitatives. En effet, toutes les associations détectées ne sont pas d'égal intérêt pour l'utilisateur et il nous a semblé qu'une recherche grossière est tout d'abord utile. Une connaissance plus approfondie de l'association peut être effectuée, dans un deuxième temps, à la demande de l'utilisateur. L'exemple précédent sur les salaires perçus par les individus exerçant une profession libérale donnerait la précision suivante : cette catégorie d'individus est sur-représentée parmi les individus touchant un salaire horaire compris entre *12,5* et *44* $. Nous avons souhaité une extraction interactive, puisque comme nous l'avons dit précédemment, toutes les associations ne sont pas d'égal intérêt pour l'utilisateur. Par ailleurs l'utilisateur a la possibilité de changer l'ensemble d'apprentissage au cours du processus d'extraction. Si nous reprenons l'exemple précédent où nous recherchions les catégories d'individus touchant de forts salaires, nous pouvons faire la même recherche non plus dans l'ensemble d'apprentissage initial mais parmi les individus exerçant une profession libérale. Ainsi, ce type de recherche a révélé que parmi les individus exerçant une profession libérale, la catégorie des individus travaillant dans le secteur de la fabrication est sur-représentée parmi les forts salaires alors que cette même association (*secteur fabrication et salaire*) n'était pas apparue dans l'ensemble d'apprentissage prenant en compte toutes les catégories socio-professionnelles.

Le papier est organisé comme suit. Dans la section deux, nous allons définir de façon précise la sémantique de ces associations quantitatives générales et atypiques. Dans la section trois, nous rappellerons la notion d'intensité d'inclination et son adaptation pour permettre l'extraction de ces associations. Dans la section quatre, nous exposerons la technique d'obtention des intervalles afin d'avoir une connaissance plus précise des associations. Nous terminerons, dans la section cinq, par la présentation du processus d'extraction interactif et par la réalisation d'expérimentations sur trois jeux de données différents. Le papier s'achèvera sur une conclusion et des perspectives de recherche.

2 Associations quantitatives générales et atypiques

Comme nous nous focalisons sur l'étude des variables quantitatives, les associations recherchées concernent toujours au moins une variable quantitative que nous nommerons variable cible. Nous verrons que l'extension aux variables catégorielles est immédiate.

Il existe deux types d'associations quantitatives, selon que l'ensemble des variables associé à la variable quantitative cible fait intervenir ou non au moins une variable quantitative. Nous appellerons le premier type d'association *Catégorielles – Quantitative* (*ou C-Q*) et le second *Quantitatives – Quantitative* (*ou Q-Q*).

2.1 Associations Catégorielles - Quantitative (*C-Q*)

Ce type d'association recherche, pour une catégorie d'individus, un comportement atypique, c'est-à-dire différent du comportement de l'ensemble des individus, pour la variable quantitative cible.

Ce type d'association est de la forme :

Catégorie d'individus : Comportement atypique de la variable quantitative cible

Nous allons rechercher les catégories d'individus qui sont sur-représentées ou au contraire sous-représentées parmi les individus vérifiant de fortes ou de faibles valeurs pour la variable quantitative cible.

Afin d'illustrer nos propos, nous allons prendre un exemple tiré de la base Wages disponible à l'adresse suivante http://lib.stat.cmu.edu/datasets. Nous nous intéressons au nombre d'années d'étude effectué par les *534* individus de cette base. La courbe de gauche de la *figure 1* montre la distribution de cette variable pour l'ensemble des individus de la base. La valeur minimale de cette variable est de *2* années, la valeur maximale est de *18* années et la moyenne est de *13,02* années. La courbe de droite de la *figure 1* représente la distribution de cette même variable mais pour les individus exerçant une profession libérale. Cette catégorie d'individus fait au minimum *6* années d'étude et va jusqu'à *18* années. La moyenne est de *15,64* années. *26%* des individus de cette base font entre *15* et *18* années d'étude (*139 individus pour un ensemble d'apprentissage de 534 individus*). On s'attend donc que pour toutes les catégories d'individus, *26%* d'entre eux font entre *15* et *18* années d'étude. Or *74,3%* des individus exerçant une profession libérale font entre *15* et *18* années d'étude (*78 individus pour une catégorie de 105 individus*). Cette catégorie d'individus est donc sur-représentée parmi les individus ayant fait de longues études.

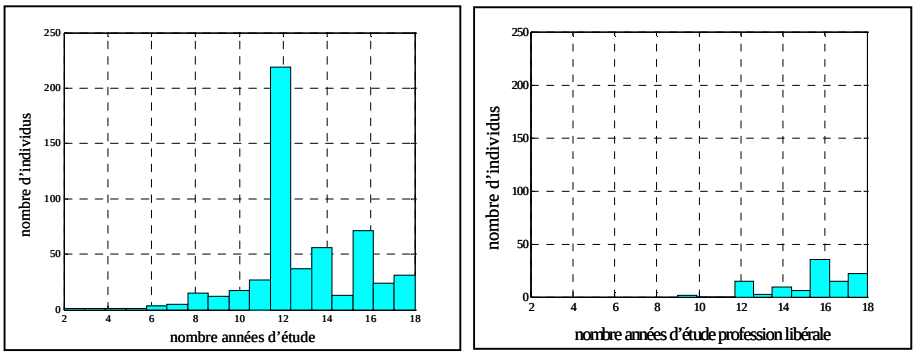


FIG. 1 – *Distribution de la variable ″Nombre d'années d'étude″ pour l'ensemble d'apprentissage (courbe de gauche) et pour les individus exerçant une profession libérale (courbe de droite).*

73,2% des individus de l'ensemble d'apprentissage font entre *6* et *14* années d'étude alors que *25,7%* des individus exerçant une profession libérale font entre *6* et *14* années d'étude. Les professions libérales sont donc sous-représentées parmi les individus ayant fait peu d'étude.

Les individus exerçant une profession libérale ont donc un comportement atypique puisqu'ils font beaucoup plus d'étude que l'ensemble des individus de la base. Cela

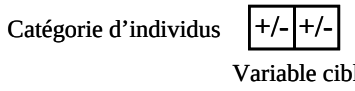
conforte notre croyance que pour exercer une profession libérale, un certain nombre d'années d'étude est indispensable.

Une sur-représentation de la variable quantitative pour la catégorie d'individus considérée sera formalisée par le symbole ″+″ et une sous-représentation par le symbole ″-″. Afin de distinguer les faibles valeurs des fortes valeurs de la variable cible, nous placerons les deux symboles ″+″ et ″-″, à gauche pour les faibles valeurs, et à droite pour les fortes valeurs de la variable cible.

La forme de l'association devient donc la suivante :

Catégorie d'individus : (+/-) Variable cible (+/-).

Nous pouvons également la représenter sous forme graphique afin de faciliter sa lecture :

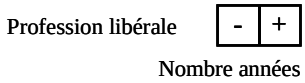


La case de gauche de la représentation graphique symbolise les faibles valeurs de la variable cible et la case de droite symbolise les fortes valeurs.

Si nous reprenons l'exemple précédent de la base *Wages*, nous avons l'association suivante :

Profession libérale : (-) Nombre d'années d'étude (+)

ou encore



Ce type d'association détecte uniquement la sous-représentativité ou la sur-représentativité pour les fortes ou les faibles valeurs de la variable cible sans rechercher les intervalles de valeurs précis pour lesquelles cette représentativité atypique a été décelée. Cette recherche d'intervalles sera développée dans la *section 4*.

2.2 Associations Quantitatives - Quantitative (*Q*-*Q*)

Ce deuxième type d'association fait intervenir dans la partie gauche de l'association, c'est-à-dire pour la catégorie d'individus, au moins une variable quantitative. Comme pour le premier type d'association *C*-*Q*, nous allons nous intéresser aux catégories d'individus possédant de faibles et de fortes valeurs pour toutes les variables quantitatives considérées en partie gauche.

La forme de ce deuxième type d'association est la suivante :

Catégorie d'individus vérifiant de faibles valeurs : (+/-) Variable cible (+/-).

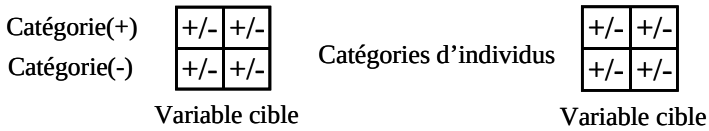
Catégorie d'individus vérifiant de fortes valeurs : (+/-) Variable cible (+/-).

que l'on peut condenser de la façon suivante :

Catégorie d'individus (-) : (+/-) Variable cible (+/-).

Catégorie d'individus (+) : (+/-) Variable cible (+/-).

ou encore représenter sous une des deux formes graphiques :



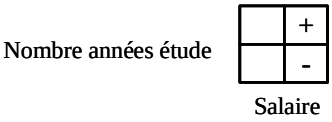
Découverte d'associations quantitatives générales et atypiques

Afin d'illustrer nos propos, nous reprenons la base de données *Wages* et étudions l'influence sur le salaire du nombre d'années d'étude effectué par les individus. Nous aimerions savoir si le salaire perçu par les individus ayant fait de courtes études et par les individus ayant fait de longues années d'étude est fort ou faible.

Comme pour les associations *C-Q*, nous allons détecter pour ces deux catégories d'individus, si le comportement de la variable cible s'écarte du comportement général. Ainsi, nous allons rechercher si ces deux catégories d'individus sont sur-représentées ou sous-représentées pour les faibles et fortes valeurs de la variable cible.

Si nous reprenons l'exemple précédent, nous avons l'association suivante :

Nombre années d'étude (+) : (-) Salaire (+)



Ainsi, les individus ayant fait de longues études, sont sur-représentées parmi les individus touchant de forts salaires et au contraire, sont sous-représentées parmi les faibles salaires.

La *figure 2* met en évidence ce comportement atypique pour les individus ayant fait de longues études. Dans un but de lisibilité du graphique, nous avons effectué une discrétisation en cinq intervalles des deux variables quantitatives, discrétisation non nécessaire pour trouver cette association générale. Ainsi parmi les *139* individus ayant fait au moins *15* années d'étude (*intervalles 4 et 5 de la variable nombre années*), *92* d'entre elles touchent au moins *9* $ par heure (*intervalles 4 et 5 de la variable salaire*), ce qui représente *66,19*%. Or dans l'ensemble d'apprentissage, *220* individus ont fait au moins *15* années d'étude (*ce qui représente 41,2% de l'ensemble d'apprentissage*), d'où la sur-représentativité détectée.

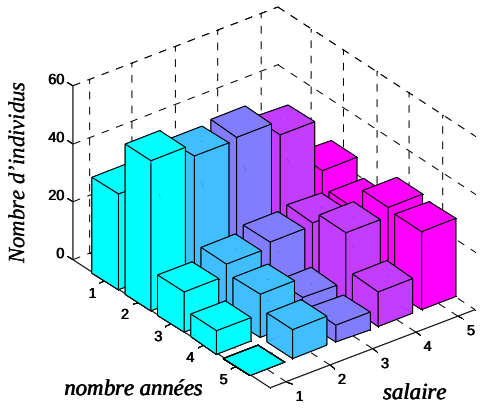


FIG. 2 – *Distribution de la variable "Nombre d'années d'étude" selon le salaire.*

3 Intensité d'inclination

Pour déceler les comportements atypiques de certaines catégories d'individus, nous allons utiliser et adapter une mesure existante, l'intensité d'inclination. Tout d'abord, nous

allons donner la sémantique de cette mesure et ensuite nous indiquerons comment nous l'avons adaptée pour déceler les associations *C-Q* et *Q-Q*.

3.1 Sémantique de la mesure

L'intensité d'inclination Guillaume (2002) est une généralisation aux variables quantitatives de l'intensité d'implication Gras (1979) (*mesure adaptée pour les variables binaires*), de l'intensité d'implication d'Annie Larher (1991) et de l'intensité de propension Lagrange (1997) (*mesures adaptées aux variables numériques à valeurs dans l'intervalle [0,1]*). Ces mesures dérivent d'une mesure de similitude, l'indice de la vraisemblance du lien Lerman (1981).

L'intensité d'implication mesure la *″petitesse″* du nombre d'individus qui violent la règle (*appelé contre-exemples*), c'est-à-dire les individus qui vérifient la prémisse et qui ne vérifient pas la conclusion. Une règle sera d'autant plus intéressante que le nombre de contre-exemples est faible.

L'intensité d'inclination évalue si le nombre des individus ayant une forte appartenance à la prémisse (*dans le cas binaire : qui vérifie la prémisse*) et une faible appartenance à la conclusion (*dans le cas binaire : qui ne vérifie pas la conclusion*) est également faible. Comme pour l'intensité d'implication, une règle sera jugée d'autant plus intéressante que ce nombre de contre-exemples est faible.

Si nous reprenons l'exemple de l'association *″Nombre d'années d'étude – Salaire ″* de la base *Wages* et que nous nous intéressions à la règle *″Nombre d'années d'étude → Salaire″*, l'intensité d'inclination va évaluer si le nombre des individus ayant une forte appartenance à *″Nombre d'années d'étude″* et une faible appartenance à *″Salaire″* est faible, c'est-à-dire si le nombre des individus ayant fait beaucoup d'étude et percevant un petit salaire est faible.

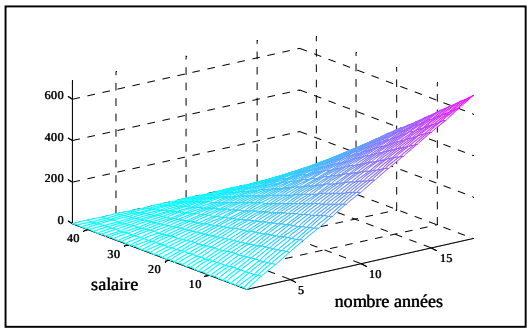


FIG. 3 – *Poids attribué aux individus dans le cas de l'association entre salaire et nombre d'années d'étude.*

Pour trouver ce nombre de contre-exemples, un poids va être attribué à chaque individu et ce poids sera d'autant plus fort qu'il se rapproche de la notion de contre-exemple. Les valeurs minimale et maximale pour la variable *″nombre d'années d'étude″* sont respectivement égales à *2* ans et *18* ans ; et les valeurs minimale et maximale pour la variable *″salaire″* sont respectivement égales à *1* $ et *44,5* $. Ainsi, un individu ayant fait *18* années d'étude et percevant un salaire horaire de *1* $ aura un poids maximum puisqu'il correspond totalement à la notion d'individus contredisant la règle ; alors qu'un individu

ayant fait *2* années d'étude et percevant un salaire de *44,5* $ aura un poids minimal égal à *0*. Les autres individus auront un poids intermédiaire compris entre le poids de ces deux individus (*premier individu : étude = 18 années et salaire = 1$, deuxième individu : étude = 2 années et salaire = 44,5$*), poids fonction de leur proximité à ces deux individus extrêmes comme le montre la *figure 3*.

Afin de savoir si ce nombre de contre-exemples est significatif, un test statistique est ensuite réalisé, test développé dans le paragraphe suivant.

3.2 Définition de la mesure

Soient X et Y deux conjonctions de respectivement p et q variables quantitatives et Ω l'ensemble d'apprentissage. On pose $X = X_1, .., X_p$ et $Y = Y_1, .., Y_q$, où $X_1, .., X_p, Y_1, .., Y_q$ sont des variables quantitatives à valeurs $x_{1_i}, .., x_{p_i}, y_{1_i}, .., y_{q_i}$ ($i \in \{1..N\}$) dans respectivement les intervalles $[x_{1_{min}} .. x_{1_{max}}], .., [x_{p_{min}} .. x_{p_{max}}], [y_{1_{min}} .. y_{1_{max}}], .., [y_{q_{min}} .. y_{q_{max}}]$.

L'intensité d'inclination mesure si le nombre des individus ne vérifiant pas fortement la règle $X \rightarrow Y$ (*c'est-à-dire le nombre des individus ayant simultanément une valeur élevée pour chacune des variables $X_1, .., X_p$ et une valeur faible pour chacune des variables $Y_1, .., Y_q$*) est significativement faible comparativement à ce que l'on obtiendrait si par hypothèse les variables X et Y étaient indépendantes.

Soient x_{min} la valeur minimale de X, x_i la valeur vérifiée par l'individu e_i ($e_i \in \Omega$) pour la variable X, y_{max} la valeur maximale de Y et y_i la valeur vérifiée par l'individu e_i pour la variable Y.

Le nombre t_0 de contre-exemples pour la règle $X \rightarrow Y$ est défini de la façon suivante :

$$t_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_{\max} - y_i) \text{ avec}$$

$$x_i = \sum_{j=1}^{p} x'_{j_i}, \; x_{\min} = \sum_{j=1}^{p} x'_{j_{\min}}, \; y_i = \sum_{k=1}^{q} y'_{k_i}, \; y_{\max} = \sum_{k=1}^{q} y'_{k_{\max}},$$

$$x'_{j_i} = \frac{x_{j_i} - \mu_{X_j}}{\sigma_{X_j}} \; (j \in \{1..p\}), \; y'_{k_i} = \frac{y_{k_i} - \mu_{Y_k}}{\sigma_{Y_k}} \; (k \in \{1..q\})$$

et où μ_{X_j}, μ_{Y_k} sont les moyennes respectivement des variables X_j ($j \in \{1, .., p\}$) et Y_k ($k \in \{1, .., q\}$) et où σ_{X_j}, σ_{Y_k} sont les écart-types respectivement de X_j et Y_k.

La variable aléatoire T, dont t_0 est une valeur observée, suit asymptotiquement la loi normale $\mathcal{N}(\mu, \sigma)$ avec $\mu = N(\mu_X - x_{min})(y_{max} - \mu_Y)$ et $\sigma^2 = N[v_X v_Y + v_Y (\mu_X - x_{min})^2 + v_X (y_{max} - \mu_Y)^2]$.

Les moyennes et variances des variables X et Y sont données par les formules suivantes :

$$\mu_X = \sum_{j=1}^{p} \mu_{X_j}, \; \mu_Y = \sum_{k=1}^{q} \mu_{Y_k}, \; v_X = \sum_{j=1}^{p} v_{X_j} + 2 \sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} \mathrm{cov}(X_j, X_{j'})$$

$$v_Y = \sum_{k=1}^{q} v_{Y_k} + 2 \sum_{k=1}^{q-1} \sum_{k'=k+1}^{q} \mathrm{cov}(Y_k, Y_{k'}) \quad \text{avec} \quad \mathrm{cov}(X_i, X_{i'}) = \mu_{X_i X_{i'}} - \mu_{X_i} \mu_{X_{i'}}.$$

Si la probabilité $Pr(T \leq t_o)$ d'avoir un nombre inférieur ou égal à t_o est élevée, nous pouvons en conclure que t_o n'est pas significativement faible car pouvant se produire assez fréquemment et par conséquent l'implication $X \rightarrow Y$ n'est pas pertinente.

Afin de mesurer cette implication de façon croissante, l'indice $\varphi(X \rightarrow Y) = 1 - F(t_o) = Pr(T > t_o)$ est retenu où F est la fonction de répartition[1] de T. Ainsi, l'implication $X \rightarrow Y$ est admissible au niveau de confiance $(1-\alpha)$ si et seulement si $Pr(T \leq t_o) \leq \alpha$ ou $Pr(T > t_o) \geq 1 - \alpha$.

L'intensité d'inclination est donc :

$$\varphi(X \rightarrow Y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{t_0}^{+\infty} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt$$

3.3 Adaptation de l'intensité d'inclination

Nous avons considéré la situation de la zone des fortes valeurs pour X et des faibles valeurs pour Y (*que nous nommerons zone* Z_1). Nous pouvons maintenant étendre notre analyse aux trois autres zones Z_2 à Z_4 (*voir figure 4*) :

zone Z_2 : fortes valeurs pour X et Y

zone Z_3 : faibles valeurs pour X et fortes valeurs pour Y

zone Z_4 : faibles valeurs pour X et Y.

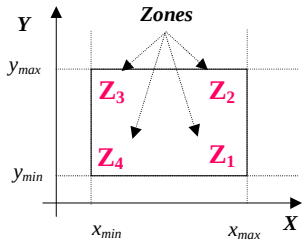


FIG. 4 – *Les différentes zones.*

L'adaptation de l'intensité d'inclination aux trois autres zones permet de rechercher la sous-représentativité d'une catégorie d'individus vérifiant X pour la variable quantitative cible Y. Nous verrons, dans un deuxième temps, comment à partir de l'extraction de la sous-représentativité, nous passons à la sur-représentativité.

Sous-représentativité. Tout d'abord, nous allons adapter le nombre de contre-exemples t_0 aux trois zones Z_2, Z_3 et Z_4. Nous savons que pour la zone Z_1, cet indice comptabilise le nombre d'individus ayant une forte appartenance à X et une faible appartenance à Y, d'où une valeur égale à $t_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_{\max} - y_i)$.

Ainsi, pour les trois autres zones, nous avons les indices suivants :

$\boldsymbol{Z_2}$: $t_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_i - y_{\min})$, (nombre d'individus ayant une forte appartenance à X et à Y).

$\boldsymbol{Z_3}$: $t_o = \sum_{i=1}^{N} (x_{\max} - x_i)(y_i - y_{\min})$, (nombre d'individus ayant une faible appartenance à X et une forte appartenance à Y).

[1] ou fonction cumulative ou encore fonction de distribution.

Découverte d'associations quantitatives générales et atypiques

$\boldsymbol{Z_4}$: $t_o = \sum_{i=1}^{N} (x_{\max} - x_i)(y_{\max} - y_i)$, (*nombre d'individus ayant une faible appartenance à X et à Y*).

Les expressions suivantes permettent alors de détecter la sous-représentativité de la catégorie d'individus vérifiant la variable *X* pour la variable quantitative cible *Y* dans les zones Z_2, Z_3 et Z_4 :

$\boldsymbol{Z_2}$: $\varphi (X \rightarrow (y_{\max} + y_{\min} - Y)) \geq 1 - \alpha$

$\boldsymbol{Z_3}$: $\varphi (Y \rightarrow X) \geq 1 - \alpha$

$\boldsymbol{Z_4}$: $\varphi ((y_{max} + y_{min} - Y) \rightarrow X) \geq 1 - \alpha$

Sur-représentativité. Nous souhaitons maintenant déterminer si nous avons plus d'individus que ce qui est attendu dans les quatre zones.

L'implication $X \rightarrow Y$ est admissible au niveau de confiance $(1-\alpha)$ si et seulement si $Pr(T \leq t_o) \leq \alpha$, ce qui nous indique que nous avons peu de contre-exemples comme cela est représenté par la *figure 5*.

Au contraire nous aurons beaucoup de contre-exemples si $Pr(T \leq t_o) \geq 1-\alpha$ et nous pourrons en déduire que nous sommes dans une zone de sur-représentativité de la catégorie d'individus vérifiant *X* pour *Y*.

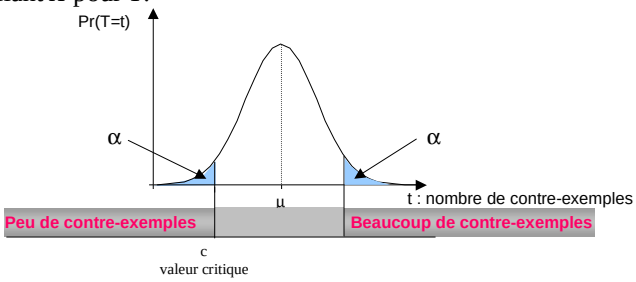


FIG. 5 – *Détermination de la sous-représentativité et de la sur-représentativité.*

Nous calculerons si les quatre régions sont sur-représentées avec les expressions suivantes :

$\boldsymbol{Z_1}$: $1 - \varphi (X \rightarrow Y) \geq 1 - \alpha$

$\boldsymbol{Z_2}$: $1 - \varphi (X \rightarrow (y_{max} + y_{min} - Y)) \geq 1 - \alpha$

$\boldsymbol{Z_3}$: $1 - \varphi (Y \rightarrow X) \geq 1 - \alpha$

$\boldsymbol{Z_4}$: $1 - \varphi ((y_{max} + y_{min} - Y) \rightarrow X) \geq 1 - \alpha$

Le *tableau 1* résume, pour chaque zone, la formule de l'intensité d'inclination à calculer pour déterminer si la catégorie d'individus vérifiant *X* est sous-représentée ou sur-représentée pour la variable *Y*.

	Sur-représentativité	***Sous-représentativité***
$\boldsymbol{Z_1}$	$1 - \varphi (X \rightarrow Y) \geq 1 - \alpha$	$\varphi (X \rightarrow Y) \geq 1 - \alpha$
$\boldsymbol{Z_2}$	$1-\varphi (X \rightarrow (y_{max}+y_{min}-Y)) \geq 1 - \alpha$	$\varphi (X \rightarrow (y_{max}+y_{min}-Y)) \geq 1 - \alpha$
$\boldsymbol{Z_3}$	$1 - \varphi (Y \rightarrow X) \geq 1 - \alpha$	$\varphi (Y \rightarrow X) \geq 1 - \alpha$
$\boldsymbol{Z_4}$	$1-\varphi ((y_{max}+y_{min}-Y) \rightarrow X) \geq 1 - \alpha$	$\varphi ((y_{max}+y_{min}-Y) \rightarrow X) \geq 1 - \alpha$

TAB. 1 – *Formules de l'intensité d'inclination.*

4 Recherche des intervalles

Dans cette section nous montrons comment l'intensité d'inclination peut être adaptée pour découvrir le comportement atypique (*sous-représentativité ou sur-représentativité*) de catégories d'individus dans des intervalles de valeurs quelconques de la variable cible.

Nous allons traiter l'extraction des intervalles dans le cas le plus général c'est-à-dire dans le cadre d'une association *Q-Q*.

Une sous-représentativité indique que nous avons moins d'individus que ce qui est attendu alors qu'une sur-représentativité indique que nous avons plus d'individus que ce qui est attendu. Nous allons donc calculer le tableau de contingence attendu à partir du tableau de contingence observé afin d'obtenir le tableau des écarts (*différence entre le tableau des effectifs observés et le tableau des effectifs attendus*) entre la situation observée et la situation attendue.

Soient les variables quantitatives X et Y prenant leurs valeurs dans respectivement les intervalles $[x_{min}, x_{max}]$ et $[y_{min}, y_{max}]$ et ayant respectivement r et s valeurs distinctes $x_1 = x_{min}$, x_2, …, $x_r = x_{max}$ et $y_1 = y_{min}$, y_2, …, $y_s = y_{max}$. Soit N le nombre total d'individus dans l'ensemble d'apprentissage Ω et soit n_{ij} ($i \in \{1,..,r\}$ *et* $j \in \{1,..,s\}$) le nombre d'individus vérifiant simultanément $X = x_i$ et $Y = y_j$. Soit le tableau de contingence T des écarts (*Tableau 2*) entre les effectifs observés n_{ij} et les effectifs attendus[2] $n_{i.}n_{.j}/N$ sous l'hypothèse d'indépendance entre les variables X et Y.

Nous avons les propriétés suivantes :

$$\sum_{k=1}^{s} n_{ik} - n_{.k}n_{i.}/N = 0 \text{ et } \sum_{k=1}^{r} n_{kj} - n_{.j}n_{k.}/N = 0$$

C'est pourquoi la somme des lignes et des colonnes du tableau de contingence des différences est nulle.

Les cellules du tableau ayant des valeurs positives indiquent que nous avons plus d'individus que ce qui est attendu et par conséquent que nous sommes dans une zone sur-représentée. Au contraire, les cellules ayant des valeurs négatives indiquent que nous avons moins d'individus que ce qui est attendu et que nous sommes dans une zone sous-représentée.

	$X = x_1$	…	$X = x_i$	…	$X = x_r$	
$Y = y_1$	n_{11}-$n_{.1}n_{1.}$/N	…	n_{i1}-$n_{.1}n_{i.}$/N	…	n_{1r}-$n_{.1}n_{r.}$/N	*0*
…	…	…	…	…	…	…
$Y = y_j$	n_{1j}-$n_{.j}n_{1.}$/N	…	n_{ij}-$n_{.j}n_{i.}$/N	…	N_{rj}-$n_{.j}n_{r.}$/N	*0*
…	…	…	…	…	…	…
$Y = y_s$	n_{1s}-$n_{.s}n_{1.}$/N	…	n_{is}-$n_{.s}n_{i.}$/N	…	n_{rs}-$n_{.s}n_{r.}$/N	*0*
	0	…	*0*	…	*0*	*0*

TAB. 2 – *Tableau de contingence T des écarts entre les effectifs observés et les effectifs attendus.*

[2] $n_{i.} = \sum_{k=1}^{s} n_{ik}$ est la distribution marginale de x_i et correspond à la distribution de $X = x_i$ sans tenir compte de Y, et $n_{.j} = \sum_{k=1}^{r} n_{kj}$ est la distribution marginale de y_j et correspond à la distribution de $Y = y_j$ sans tenir compte de X.

Afin d'illustrer nos propos, nous allons reprendre l'association entre le nombre d'années d'étude et le salaire. Le *tableau 3* donne le tableau de contingence obtenu à partir de l'ensemble d'apprentissage. Ainsi, nous n'avons pas d'individu ayant un salaire inférieur strictement à *5* $ et qui a poursuivi *17* ou *18* années d'étude.

Le *tableau 4* nous restitue le tableau de contingence *T* des écarts entre les effectifs observés et les effectifs théoriques. Pour les variables *Salaire = [0 ; 5[* et *Nombre années = [17 ; 18]*, l'effectif attendu est de *11* (*55×107 / 534*) individus et pour *Salaire = [12,5 ;44]* et *Nombre années = [15 ;16]*, l'effectif attendu est de *17* (*84×107 / 534)* individus.

	Sal <5	***5≤Sal<6,67***	***6,67≤Sal<9***	***9≤Sal<12,5***	***12,5≤Sal***	***Total***
2≤Edu≤11	33	16	17	11	6	*83*
Edu = 12	52	48	49	44	26	*219*
13≤Edu≤14	14	18	20	21	20	*93*
15≤Edu≤16	8	15	8	25	28	*84*
17≤Edu≤18	0	10	6	12	27	*55*
Total	*107*	*107*	*100*	*113*	*107*	*534*

TAB. 3 – *Tableau de contingence.*

	Sal <5	***5≤Sal<6,67***	***6,67≤Sal<9***	***9≤Sal<12,5***	***12,5≤Sal***
2≤Edu≤11	16	-1	1	-7	-11
Edu = 12	8	4	8	-2	-18
13≤Edu≤14	-5	-1	3	1	1
15≤Edu≤16	-9	-2	-8	7	11
17≤Edu≤18	-11	-1	-4	0	16

TAB. 4 – *Tableau de contingence T des écarts (les totaux des lignes et colonnes sont nuls).*

Ainsi, le *tableau 4* nous révèle que pour *Salaire = [0, 5[* et *Nombre années = [17 ; 18]*, nous avons *11* individus en moins par rapport à ce qui est attendu sous l'hypothèse que *X* et *Y* soient indépendantes ($n_{ij} - n_{i.}n_{.j}/N = 0 - 11 = -11$). Au contraire, pour *Salaire = [12,5 ;44]* et *Nombre années = [15 ;16]*, nous avons *11* individus en plus par rapport à ce qui est attendu ($n_{ij} - n_{i.}n_{.j}/N = 28 - 17 = 11$).

Comme nous recherchons les plus grandes zones rectangulaires de valeurs positives et de valeurs négatives, les intervalles mis en jeu sont donc les suivants :

Nombre années = [*13 ; 18*] et *Salaire* = [*9 ; 44*] pour la sur-représentation,

Nombre années = [*15 ; 18*] et *Salaire* = [*0 ;9*[pour la sous-représentation.

Les deux autres zones de valeurs positives (*Nombre années* = [*2 ; 12*] et *Salaire* = [*0 ; 5*[) et de valeurs négatives (*Nombre années* = [*2 ; 12*] et *Salaire* = [*9 ; 44*]) n'ont pas été jugées significatives par l'intensité d'inclination (*intensité minimale de 0,95*).

Ainsi, les individus ayant fait des études de plus de *13* années sont sur-représentés parmi les individus gagnant au moins *9* $ et les individus ayant fait plus de *15* ans d'étude sont sous-représentés parmi les individus gagnant au plus *9* $.

5 Processus d'extraction et expérimentations

La recherche des associations en présence d'une variable quantitative est un processus d'extraction par niveaux, interactif et centré utilisateur.

Le processus démarre tout naturellement par la sélection de la variable quantitative cible. Nous nous sommes focalisés sur les variables quantitatives comme variable cible mais la même démarche peut s'appliquer aux variables catégorielles puisque ces variables prennent deux valeurs : *0* (*absence de la caractéristique*) et *1* (*présence de la caractéristique*).

La *figure 6*, illustre le processus d'extraction par niveaux. L'extraction commence par la recherche des variables *X* dont l'association avec la variable cible *Y* est intéressante. Nous dirons que c'est le *niveau 1* de l'extraction ; il correspond à la recherche des *1*-variables intéressantes (*point 1 de la figure 6*).

Une fois ces variables extraites, l'utilisateur peut demander :

- soit de voir toutes les associations avec la variable *Y* sous forme textuelle ou sous forme graphique (*point 2 de la figure 6*) après avoir éventuellement appliqué un filtre sur les variables,
- soit se focaliser sur une variable *X* particulière en demandant de restituer l'association avec la variable cible (*point 7 de la figure 6*),
- soit de lancer l'extraction des associations *X*-*Y* où *X* se compose de deux variables (*point 3 de la figure 6*). C'est le *niveau 2* de l'extraction.

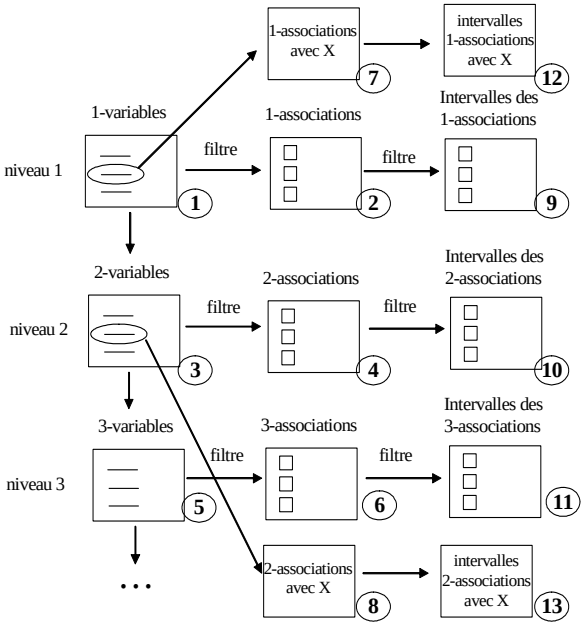


FIG. 6 – *Processus d'extraction des associations quantitatives.*

Si l'utilisateur demande l'extraction du *niveau 2*, le système va, comme pour le *niveau 1*, restituer tous les couples de variables où une association avec *Y* a été jugée intéressante (*point 3 de la figure 6*). L'utilisateur a, comme pour le niveau précédent, plusieurs options à sa disposition :

- soit visualiser toutes les associations avec la variable *Y* sous forme textuelle ou sous forme graphique (*point 4 de la figure 6*) après avoir éventuellement appliqué un filtre sur les variables,

Découverte d'associations quantitatives générales et atypiques

- soit s'intéresser à un couple de variables en demandant la restitution de l'association avec la variable cible (*point 8 de la figure 6*),
- soit lancer l'extraction du *niveau 3*, c'est-à-dire des associations *X*-*Y* où *X* se compose de trois variables (*point 5 de la figure 6*).

L'utilisateur peut à partir de toutes les *n*-associations (*points 2, 4 et 6 de la figure 6*) demander l'extraction des intervalles après avoir éventuellement appliqué un filtre sur les variables (*points 9, 10 et 11 de la figure 6*).

Il peut également se focaliser sur une association X_1 - *Y* et effectuer les mêmes requêtes que dans le cas général, à savoir :

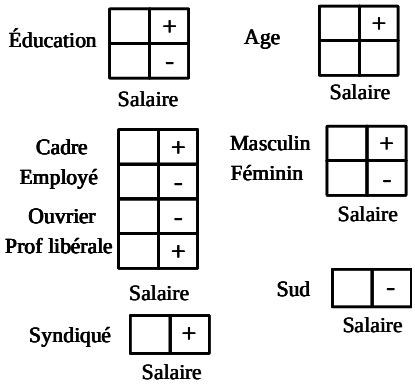
- extraire les intervalles de cette association (*points 12 et 13 de la figure 6*),
- extraire les associations de niveau supérieur impliquant obligatoirement X_1 dans l'association.

L'algorithme a été implémenté avec Matlab essentiellement pour des raisons de facilité de développement. Il s'agissait de valider l'opérationnalité du processus et d'étudier sa capacité d'analyse sur quelques situations typiques. L'étude du passage à l'échelle sera effectuée avec un prototype plus efficient. Ce prototype, en cours de réalisation par Guillochon (2007), étend le logiciel libre d'extraction de connaissances WEKA (***W****aikato* ***E****nvironnement for* ***K****nowledge* ***A****nalysis*) développé par l'université de Waikato en Nouvelle-Zélande et écrit en Java. Toutes les informations concernant le téléchargement et la documentation de WEKA sont disponible à l'adresse suivante : http://weka.sourceforge.net/wiki/index.php/Main_Page.

La recherche de ce type d'association a été effectuée sur trois ensembles de données: *Wages* et *Abalone* de la base UCI (*Murphy et Aha (1995) et Bay (2001b)*) et les *Iris* de Fisher (*Fisher (1936)*). Toutes les associations extraites sont admissibles avec un niveau de confiance de *0,95*.

La base de données *Wages* se compose de *534* enregistrements décrits par *11* variables dont *4* variables quantitatives : Éducation (*nombre d'années d'étude*), Expérience (*nombre d'années d'expérience professionnelle*), Salaire (*dollars par heure*) et Age. Les variables catégorielles sont : Région (*Nord et Sud*), Sexe, Syndique, Race (*Hispanique, Blanche et Autre*), Emploi (*Cadre, Vendeur, Employé, Ouvrier, Profession libérale et Autre*), Secteur (*Fabrication, Construction et Autre*) et Marié.

Pour cette base, nous nous sommes focalisés sur la variable *"Salaire"* et la découverte des *1*-associations (*point 2 de la figure 6*). Ces dernières nous révèlent que les catégories d'individus qui sont sur-représentées parmi ceux qui perçoivent un fort salaire sont les suivantes :



- les individus ayant fait de nombreuses années d'étude,
- les cadres (*12,70* $) et les professions libérales (*11,95* $),
- les individus d'une certaine maturité,
- les hommes (*9,99* $),
- les individus syndiquées (*10,80* $).

De la même façon, nous apprenons que les catégories d'individus suivantes sont sous-représentées parmi les individus touchant un fort salaire :

- les individus ayant fait peu d'étude,
- les employés (*7,42* $) et les ouvriers (*6,54* $),
- les femmes (*7,88* $),
- les individus vivant au sud des États-Unis (*7,90* $).

Afin de conforter les résultats obtenus, nous avons indiqué la moyenne du salaire pour toutes ces catégories d'individus sachant que la moyenne pour l'ensemble d'apprentissage est de *9,02* $.

Catégorie d'individus	**Représentativité**	**Intervalle de la variable Salaire**
Éducation=[15,18]	+	[9, 44,5]
Éducation=[2,12]	-	[9, 44,5]
Cadre	+	[10 ; 44,5]
Profession libérale	+	[10 ; 44,5]
Age=[30, 64]	+	[14 ; 44,5]
Masculin	+	[14 ; 44,5]
Syndiqué	+	[8,49 ; 44,5]
Employé	-	[14 ; 44,5]
Ouvrier	-	[9,1 ; 44,5]
Féminin	-	[5,5 ; 44,5]
Sud	-	[8,49 ; 44,5]

TAB. 5 – *Affinage des associations.*

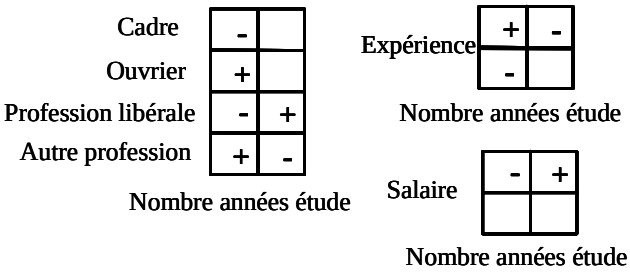
Le *tableau 5* donne les résultats d'une demande de précision des associations précédentes. La première colonne donne la catégorie d'individus concernée, la deuxième indique si la représentativité est forte (*symbolisée par le caractère ″+″*) ou faible (*symbolisée par le caractère ″-″*) et la troisième colonne restitue l'intervalle de la variable *″salaire″*.

Ainsi, la *ligne 1 du tableau 5* nous indique que les individus ayant fait au moins *15* années d'étude sont sur-représentés parmi les individus gagnant au moins *9* $ et la *ligne 6 du tableau 5* nous révèle que les hommes sont sur-représentés parmi les individus gagnant plus de *14* $.

Si nous nous focalisons maintenant sur la variable *″Profession libérale″* et que nous lançons les extractions de niveau supérieur dans ce nouvel ensemble d'apprentissage, nous apprenons que les individus travaillant dans le secteur de la fabrication sont sur-représentés parmi les individus gagnant un fort salaire et plus particulièrement parmi les individus gagnant au moins *9,5* $. L'association *″Fabrication - Salaire″* n'était pas apparue auparavant.

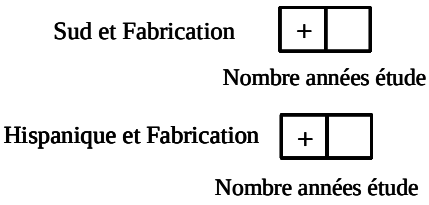
Découverte d'associations quantitatives générales et atypiques

Si nous considérons la variable cible *″Nombre d'années d'étude″*, les associations extraites nous fournissent les résultats suivants :



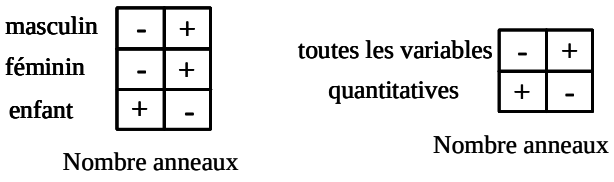
Ainsi, nous apprenons que :
- les cadres et les professions libérales sont sous-représentés parmi les individus ayant fait de courtes études.
- les professions libérales sont sur-représentées parmi les individus ayant fait de longues études alors que les cadres ne sont pas sur-représentés.
- les ouvriers et les individus exerçant une autre profession (*en dehors des cadres, employés, ouvriers, vendeurs et professions libérales*) sont sur-représentés parmi ceux qui ont fait de courtes études, seules les autres professions sont sous-représentées parmi ceux qui ont fait de longues études.
- les individus ayant une forte expérience professionnelle sont sur-représentées parmi ceux qui ont fait de courtes études ;
- les individus ayant fait de longues études sont sur-représentés parmi ceux qui gagnent un fort salaire.

En lançant les extractions de niveau supérieur, trois nouvelles variables (*Sud, Fabrication et Hispanique*) sont apparues alors qu'elles étaient non présentes au *niveau 1*.



Ainsi, les individus travaillant dans le secteur de la fabrication et soit habitant le sud des États-Unis, soit étant hispanique sont sur-représentés parmi ceux qui ont fait peu d'années d'étude.

La base de données *Abalone* se compose de *4 177* coquillages décrits par *9* variables dont *8* variables quantitatives : Longueur, Diamètre, Taille, Poids total, Poids de la chair, Poids des intestins, Poids de la carapace et Nombre d'anneaux. La dernière variable nous renseigne sur le Sexe du coquillage (*Masculin, Féminin et Enfant*). Nous nous sommes focalisés sur la variable quantitative *″Nombre d'anneaux″* et nous avons obtenu les associations suivantes :



Plus le nombre d'anneaux est important, plus la longueur, le diamètre, la taille et les poids sont importants. Pour finir, seuls les enfants sont sur-représentés parmi ceux qui possèdent peu d'anneaux.

Afin de montrer que nous pouvons faire ce même type de recherche sur des données catégorielles, nous avons choisi la base des Iris et nous nous sommes focalisés sur les trois catégories de fleurs : les iris setosa, les iris virginica et les iris versicolor. L'intérêt de cette base est d'essayer de prédire ces trois catégories de fleurs en fonction de la longueur et de la largeur des pétales et des sépales, qui sont les variables explicatives de cette base.
Nous obtenons les associations suivantes :

setosa	+	-
virginica	-	+
versicolor		

Longueur pétale

setosa	+	-
virginica	-	+
versicolor		

Largeur pétale

setosa	+	-
virginica	-	+
versicolor		

Longueur sépale

setosa	-	+
virginica		
versicolor	+	-

Largeur sépale

Ainsi, nous vérifions que la longueur des pétales pour les iris setosa est plutôt faible contrairement aux iris virginica qui ont une valeur plutôt élevée. Nous vérifions également que la largeur des pétales est plutôt élevée pour les virginica.

6 Conclusion et perspectives

Nous avons proposé un nouveau type d'association permettant de caractériser parmi une population d'individus les catégories qui sont sur-représentées ou sous-représentées par rapport au comportement normal de la population relativement à une variable quantitative cible. Nous avons proposé également un processus général d'extraction par niveaux de ce type d'association. Dans un premier temps il est possible de rechercher ce comportement atypique pour les faibles ou les fortes valeurs de la variable cible. On obtient alors une vision globale de l'association. Dans un deuxième temps on peut rechercher plus précisément les intervalles de valeurs de la variable cible pour lesquels ce comportement se manifeste. Le processus d'extraction a été implémenté en Matlab et expérimenté sur des jeux de données de faible taille. Nous avons pu ainsi validé complètement les possibilités d'analyse offertes par cette approche. Il apparaît le processus en deux temps est particulièrement apte à faire découvrir des comportements inattendus. Pour permettre de tester les problèmes de performances sur des jeux de données de grande taille, nous avons entrepris la réalisation d'un prototype plus efficient. Ce nouveau type de connaissance présente l'avantage d'éviter des calculs préliminaires coûteux comme une couverture minimale ou des motifs fréquents. Il évite également des étapes de discrétisation et de codage disjonctif complet des variables quantitatives. La recherche d'une précision est

entreprise uniquement pour les situations qui présentent un intérêt pour l'utilisateur. Cette étude a été menée dans le cas où la distribution de la variable cible est normale. Il serait possible de l'étendre à d'autres types de distribution. Dans les applications réelles, il ne suffit pas de déterminer les catégories qui ont un comportement atypique. Il faut aussi déterminer celles qui contribuent le plus à un objectif (*profit*) ou qui contribuent le moins à une charge (*coût*). Dans cette optique il serait intéressant de caractériser les catégories extraites par une mesure supplémentaire représentative de cet objectif ou de cette charge.

Références

Agrawal R., Imielinski T. and Swami A. (1993), Mining Association Rules between Sets of Items in Large Databases, *ACM-SIGMOD International Conference on Management of Data (SIGMOD'93)*, Washington, D.C., ACM Press, 207-216.

Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.I. (1996), Fast Discovery of Association Rules, In Fayyad U.M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R. eds., *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press., 307-328.

Auman Y., Lindell Y. (1999), A Statistical Theory for Quantitative Association Rules, *5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 99), 261-270.

Antonie M. and Zaane O. (2004), *Mining Positive and Negative Association Rules : An Approach for Confined Rules*, Technical Report TR04-07, Dept. of Computing Science, University of Alberta.

Bay S.D. (2001a), *Multivariate Discretization for Set Mining*, Knowledge and Information Systems, vol.3, n°4, 491-512.

Bay S.D. (2001b), The UCI KDD archive. [http://kdd.ics.uci.edu] Irvine, CA : University of California, Department of Information and Computer Science.

Brin, S. Rastogi, R. and Shim, K. (2005), Mining Optimized Gain Rules for Numeric Attributes, *IEEE transactions on Knowledge and Data Engineering*, 324-338.

Fisher R. (1936), The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics 7*.

Fukuda T., Morishita S., and Tokuyama T. (1996), Mining Optimized Association Rules for Numeric Attributes, *ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*.

Georgii E., Richter L., Rückert U. and Kramer S. (2005), Analyzing microarray data using Quantitative Association Rules, *Bioinformatics*, Oxford University Press, Vol.21, Suppl.2, 123-129.

Guillaume S. (2002), Discovery of Ordinal Association Rules, *6th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD'02), 322-327, Taipei, Taiwan.

Gras R. (1979), *Contribution à l'Etude Expérimentale et à l'Analyse de certaines Acquisitions Cognitives et de certains Objectifs Didactiques en Mathématiques*, Thèse d'Etat, Université de Rennes I.

Guillochon F. (2007), *Extension du Logiciel d'Extraction de Connaissances WEKA*, rapport de stage de 2ème année I.U.T. département informatique de Clermont-Ferrand.

Kuok, C.M. Fu, A., and Wong, M.H. (1998), Mining Fuzzy Association Rules in Databases, *ACM SIGMOD Record*, 41-46.

Lagrange J.B. (1997), Analyse Implicative d'un Ensemble de Variables Numériques; Application au Traitement d'un Questionnaire à Réponses Modales Ordonnées, *Revue de Statistique Appliquée*, I.H.P., Paris.

Larher A. (1991), *Implication Statistique et Applications à l'Analyse de Démarches de Preuve Mathématique*, Thèse d'Etat, Rennes.

Lee H-J., Park W-H. and Park D-S (2004), An Efficient Method for Quantitative Association Rules to Raise Reliance of Data, *Advances Web Technologies and Applications*, Lecture Notes in Computer Science, Springer Berlin, 506-512, ISBN 978-3-540-21371-0.

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

Ludl M.C. and Widmer G. (2000), Relative Unsupervised Discretization for Association Rule Mining, Proc. 4th European Conference Principles and Practice of Knowledge Discovery in Databases, 148-158.

Mannila H., Toivonen H. and Verkamo A.I. (1994), Efficient algorithms for Discovering Association Rules. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases*, 181-192, Seattle, Washington.

Mehta S. and Parthasarathy S. (2005), *Toward Unsupervised Correlation Preserving Discretization*, IEEE Transactions on Knowledge and Data Engineering, vol.17, n°9, 1174-1185.

Miller R.J., Yang Y. (1997), Association Rules over Interval Data, *ACM SIGMOD International Conference Management of Data*, 452-461, Tucson, AZ.

Murphy P.M. and Aha D.W. (1995), *UCI Repository of Machine Learning Databases*. Machine-readable collection, Dept of Information and Computer Science, University of California, Irvine.

Rückert U., Richter L. and Kramer S. (2004), Quantitative Association Rules Based on Half-Spaces : An Optimization Approach, In Proceedings of the *4th IEEE International Conference on Data Mining* (ICDM 04), 507-510.

Salleb-Aouissi A., Vrain C. and Nortet C. (2007), QuantMiner : a Genetic Algorithm for Mining Quantitative Association Rules, IJCAI, 1035-1040.

Srikant, R., Agrawal, R. (1996) Mining Quantitative Association Rules in Large Relational Tables, *ACM-SIGMOD International Conference Management of Data*, Montréal, Canada.

Subramanyam R.B.V. and Goswami A. (2006), Mining Fuzzy Quantitative Association Rules, *Expert Systems*, Vol. 23, N°4, 212-225.

Tong Q., Yan B. and Zhou Y. (2005), Mining Quantitative Association Rules on Overlapped Intervals, *Advanced Data Mining and Applications*, Springer Berlin,Vol. 3584, 43-50, ISBN 978-3-540-27894-8.

Yuan X., Buckles B.P., Yuan Z. and Zhang J. (2002), Mining Negative Association Rules, *Computers and Communications* (ISCC'02), 623-628.

Webb G.I. (2001), Discovering associations with numeric variables, *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 383-388, ACM Press.

Witten I.H. and Frank E. (2005), *Data Mining, practical machine learning tools and techniques with Java implementations*, Morgan Kauffman, ISBN 0-12-088407-0.

Wu X., Zhang C., and Zhang S. (2004), *Efficient Mining of both Positive and Negative Association Rules*, ACM Press, NY, USA, 381-405.

Zhang W. (1999), Mining Fuzzy Quantitative Association Rules, *11th IEEE International Conference on Tools with Artificial Intelligence*.

Summary

In this paper, we introduce a new kind of association composed of at least one quantitative variable. Our purpose is to look for in a population of individuals, the categories which deviate significantly from the normal behavior of this population. More exactly we look for categories of individuals which are overrepresented or on the contrary sub-represented for the high or the low values of a target variable. To characterize the association, we used and spread an existing measure, the intensity of inclination. As all the categories (*or associations of variables*) are not of equal interest for the user, this type of knowledge allows, at first, to have a global vision of the association. In a second time it is possible to look for the intervals for which deviations exist. Contingency tables between observed and expected situations allow these intervals to be found.