Détection de groupes atypiques pour une variable cible quantitative

Sylvie Guillaume*, Florian Guillochon*, Michel Schneider*

* Laboratoire LIMOS, UMR 6158 CNRS, Université Blaise Pascal
Complexe scientifique des Cézeaux, 63177 Aubière Cedex - France
sylvie.guillaume@isima.fr, flo.guillochon@orange.fr, michel.schneider@isima.fr

Résumé. Une tâche importante en analyse des données est la compréhension de comportements inattendus ou atypiques de groupes d'individus. Quelles sont les catégories d'individus qui gagnent de particulièrement forts salaires ou au contraire, quelles sont celles qui ont de très faibles salaires ? Nous présentons le problème d'extraction de tels groupes atypiques vis-à-vis d'une variable cible quantitative, comme par exemple la variable ″salaire″, et plus particulièrement pour les faibles et fortes valeurs d'un intervalle déterminé par l'utilisateur. Il s'agit donc de rechercher des conjonctions de variables dont la distribution diffère significativement de celle de l'ensemble d'apprentissage pour les faibles et fortes valeurs de l'intervalle de cette variable cible. Une adaptation d'une mesure statistique existante, l'intensité d'inclination, nous permet de découvrir de tels groupes atypiques. Cette mesure nous libère de l'étape de transformation des variables quantitatives, à savoir l'étape de discrétisation suivie d'un codage disjonctif complet. Nous proposons donc un algorithme d'extraction de tels groupes avec des règles d'élagage pour réduire la complexité du problème. Cet algorithme a été développé et intégré au logiciel d'extraction de connaissances WEKA. Nous terminons par un exemple d'extraction sur la base de données IPUMS du bureau de recensement américain.

1 Introduction

Un problème important en analyse des données est la compréhension de comportements inattendus ou atypiques de groupes d'individus. Quelles sont les catégories d'individus qui gagnent de particulièrement forts salaires ou au contraire, quelles sont celles qui ont de très faibles salaires ?

Notre but est de détecter automatiquement tous les groupes d'individus ayant un comportement différent de celui de l'ensemble d'apprentissage pour une variable quantitative donnée et plus particulièrement pour les faibles et les fortes valeurs d'un intervalle déterminé par l'utilisateur. Nous recherchons donc les motifs ou conjonctions de variables dont la distribution diffère significativement de celle de l'ensemble d'apprentissage pour les faibles et fortes valeurs de l'intervalle de cette variable cible.

Un domaine d'étude proche de notre travail est l'extraction des règles d'association (Agrawal et al., 1996). Les règles d'association sont des relations entre les variables de la forme $X \rightarrow Y$. Dans le cadre des bases de données transactionnelles, X et Y sont des items (*articles*) comme par exemple, `cidre` ou `crêpes` et dans le cadre des bases de données relationnelles, X et Y sont des paires d'attribut-valeur comme par exemple `salaire>20K€` ou `profession="cadre"`. Pour extraire des règles d'association, on doit tout d'abord rechercher les motifs (*ou conjonctions de variables*) ayant un support (*ou taux de couverture*) supérieur à une certaine valeur définie par l'utilisateur. Ces motifs vérifiant ce support minimum sont dits des motifs fréquents. Ensuite, à partir des motifs fréquents, on recherche les règles dont la confiance (*ou probabilité conditionnelle*) est supérieure à un autre seuil défini par l'utilisateur. Pour extraire ces groupes atypiques, nous recherchons non seulement les motifs fréquents (*pour obtenir des groupes d'individus dignes d'intérêt*) mais également les motifs dont le support est surprenant, c'est-à-dire étonnamment faible ou au contraire étonnamment élevé par rapport à ce qui est attendu. Cette extraction de motifs surprenants est réalisée grâce à l'adaptation d'une mesure statistique existante, l'intensité d'inclination (Guillaume, 2002). Cette mesure a également l'avantage de nous libérer de l'étape de transformation des variables quantitatives, à savoir l'étape de discrétisation suivie d'un codage disjonctif complet, puisqu'elle affecte un poids plus ou moins important aux individus répondant au critère recherché, comme par exemple les individus gagnant un fort salaire. (Srikant et Agrawal, 1996) effectuent une telle transformation pour extraire des règles d'association quantitatives en réalisant une discrétisation automatique capable de maîtriser la perte d'information engendrée par cette transformation. Kuok et al. (1998), Zhang (1999) et Subramanyam et Goswami (2006) utilisent la technique des ensembles flous pour rechercher de telles règles. Notre approche s'apparente à la technique des ensembles flous avec l'attribution d'un poids aux individus, poids qui peut être rapproché d'un degré d'appartenance. La recherche des bons intervalles est un problème majeur pour extraire des règles pertinentes en présence de variables quantitatives. Ludl et Widmer (2000) ainsi que Bay (2001) ont montré qu'une discrétisation de ces variables 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. D'autres auteurs ont opté pour l'optimisation des 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*). Salleb-Aouissi et al. (2007) recherchent des règles d'association quantitatives en utilisant un algorithme génétique qui découvre de façon dynamique les meilleurs intervalles des variables quantitatives en optimisant le support et la confiance d'une règle. D'autres auteurs ont préféré rechercher de nouveaux types de règles, adaptés à la spécificité des variables quantitatives. Auman et Lindell (1999) proposent des règles qui s'appuient sur la distribution des valeurs des variables quantitatives. Ainsi, une règle est jugée intéressante pour ces auteurs si la catégorie d'individus vérifiant la prémisse a une moyenne pour la variable quantitative cible significativement différente du reste de l'ensemble d'apprentissage. Rückert et al. (2004) recherchent des règles à partir d'hyperplans et obtiennent, non plus des règles à partir des motifs (*règles à partir d'hyperrectangles*), 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.

Cet article s'organise de la façon suivante. La *section 2* définit la notion de groupes atypiques et la *section 3* présente la mesure utilisée pour extraire de tels groupes. La *section 4*

expose deux critères qui vont permettre d'obtenir des groupes atypiques réellement intéressants, critères qui seront ensuite utilisés dans la *section 5* lors de la présentation de l'algorithme pour réduire la complexité du problème. La *section 6* évalue l'approche retenue sur une base de données standard : la base IPUMS du bureau de recensement américain. Nous terminons par une conclusion et des perspectives.

2 Groupes atypiques

Dans cette section, nous allons définir la notion de groupes atypiques, c'est-à-dire des groupes qui sont étonnamment sur-représentés ou au contraire étonnamment sous-représentés dans une zone de la variable cible définie par l'utilisateur.

Il nous a semblé intéressant de laisser à l'utilisateur la possibilité d'étudier cette variable quantitative cible dans une zone ou sur un intervalle particulier pour plusieurs raisons. Tout d'abord, cela permet d'éliminer les valeurs non définies pour cette variable en raison d'erreurs de saisie ou de valeurs mises volontairement en dehors des valeurs permises pour indiquer que la variable ne peut être renseignée pour l'individu étudié. Ensuite, cela permet d'éliminer les individus exceptionnels qui ont soit de très fortes valeurs soit de très faibles valeurs pour la variable cible et qui, de par leur présence, risque de biaiser les résultats. Par exemple, lors d'une étude sur les salaires, il est préférable d'éliminer les individus ayant une très grosse fortune comme par exemple l'auteur des récits d'un célèbre petit sorcier. Pour finir, cela permet de faire une étude sur un segment particulier de l'ensemble d'apprentissage, comme par exemple, une étude sur les personnes percevant un salaire moyen.

Notre étude se focalise donc sur une variable cible quantitative Z et plus particulièrement sur la zone d'intérêt $Z = [z_1, z_2]$.

Soit M un motif ou une conjonction de variables. Nous travaillons sur deux types de motifs : les motifs qualitatifs X, c'est-à-dire les motifs composés uniquement de variables qualitatives ayant subi un codage disjonctif complet (*i.e. couples de variable qualitative-valeur comme par exemple profession=″cadre″*) et les motifs quantitatifs XY, c'est-à-dire les motifs qualitatifs X auquel on a ajouté une variable quantitative Y, variable quantitative sur laquelle aucune transformation n'a été opérée. Nous assimilons également les variables quantitatives Y (*variables Y n'ayant subi aucune transformation*) à des motifs quantitatifs. Nous nous limitons à une variable quantitative par motif quantitatif car l'interprétation des résultats est plus délicate en présence de plusieurs variables quantitatives. Dans cet article, nous nous intéressons aux motifs M combinés à notre variable cible $Z = [z_1, z_2]$. Nous appellerons association ciblée, la conjonction d'un motif M avec notre zone d'intérêt $Z = [z_1, z_2]$. Par simplification, nous noterons ZM cette association ciblée. Lorsque le motif $M=X$ est qualitatif, nous parlerons d'association ciblée qualitative ZX, et lorsque le motif $M=XY$ est quantitatif, nous parlerons d'association ciblée quantitative ZXY.

Dans le calcul du support d'un motif M (Agrawal et al., 1996), le même poids est attribué à chacun des individus vérifiant le motif. Nous nous intéressons aux associations ciblées ZM, dans lesquelles aucune transformation n'a été effectuée sur Z. Nous souhaitons connaître les catégories d'individus qui ont un support étonnamment élevé, ou au contraire étonnamment faible, pour les deux zones de l'intervalle de Z : les faibles et les fortes valeurs de cet intervalle. Pour cela, nous attribuons un poids aux individus, poids qui sera d'autant plus important que l'individu est en adéquation avec le critère recherché, comme par exemple ″toucher

un fort salaire", donc attribution d'un poids plus important aux individus ayant de fortes valeurs pour la variable *"salaire"*. Nous sommes donc amenés à définir deux nouveaux supports : le support positif qui se focalise sur les fortes valeurs de l'intervalle de *Z* et le support négatif qui se focalise sur les faibles valeurs de l'intervalle de *Z*.

Soient Ω l'ensemble d'apprentissage et e_i un individu de l'ensemble d'apprentissage. Soient $C = (Z = [z_1, z_2])_{e_i \in \Omega}$ l'ensemble des individus ayant une valeur pour la variable *Z* comprise entre z_1 et z_2 et $(X)_{e_i \in \Omega}$ l'ensemble des individus vérifiant le motif *X*. Soient z_i la valeur prise par l'individu e_i pour la variable cible *Z* et x_i la valeur prise par l'individu e_i pour la variable qualitative *X* (x_i = *1 si l'individu* e_i *vérifie le motif X*, x_i = *0 sinon*).

Définition 1. Le **support positif** d'une association ciblée qualitative *ZX*, que nous noterons $\text{supAbs}(ZX,+)$, est le support absolu pondéré du motif *X* dans l'ensemble *C* où l'on accorde un poids plus important aux individus ayant une valeur de *Z* proche de z_2.

$$\text{supAbs}(ZX,+) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_i\text{-}z_1) = \sum_{e_i \in C} (z_i - z_1) x_i$$

Définition 2. Le **support négatif** d'une association ciblée qualitative *ZX*, que nous noterons $\text{supAbs}(ZX,-)$, est le support absolu pondéré du motif *X* dans l'ensemble *C* où l'on accorde un poids plus important aux individus ayant une valeur de *Z* proche de z_1.

$$\text{supAbs}(ZX,-) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_2\text{-}z_i) = \sum_{e_i \in C} (z_2 - z_i) x_i$$

Dans le cas d'une association ciblée quantitative *ZXY*, nous avons en plus la connaissance des valeurs prises par les individus de l'ensemble $(X)_{e_i \in \Omega} \cap C$ pour cette variable quantitative *Y*. Comme nous ne souhaitons pas effectuer de transformation sur les variables quantitatives, nous allons également nous intéresser aux zones de faibles et fortes valeurs prises par cette variable *Y*. Les notions de supports positifs et de supports négatifs donnent lieu respectivement à deux autres notions : le support (*positif ou négatif*) pour les faibles valeurs de *Y* et le support (*positif ou négatif*) pour les fortes valeurs de *Y*. Contrairement à la variable cible *Z* où l'utilisateur définit une zone d'intérêt [z_1, z_2], pour les autres variables quantitatives *Y* associées à *Z*, nous ne laissons pas la possibilité à l'utilisateur de faire des choix ou des restrictions.

Définition 3. Le **support positif** d'une association ciblée quantitative *ZXY* est le support absolu pondéré du motif *XY* dans l'ensemble *C* où l'on accorde un poids plus important à la fois aux individus ayant une valeur de *Z* proche de z_2 et

- aux individus ayant une forte valeur pour *Y* (*d'où l'association ciblée ZXY+*),
- aux individus ayant une faible valeur pour *Y* (*d'où l'association ciblée ZXY-*).

$$\text{supAbs}\ (ZXY+,+) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_i - z_1)(y_i - y_{\min})$$

$$\text{supAbs}\ (ZXY-,+) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_i - z_1)(y_{\max} - y_i)$$

avec y_i la valeur prise par l'individu e_i pour la variable *Y*.

Définition 4. Le **support négatif** d'une association ciblée quantitative *ZXY* est le support absolu pondéré du motif *XY* dans l'ensemble *C* où l'on accorde un poids plus important à la fois aux individus ayant une valeur de *Z* proche de z_1 et

- aux individus ayant une forte valeur pour *Y* (*association ciblée ZXY+*),
- aux individus ayant une faible valeur pour *Y* (*association ciblée ZXY-*).

$$\text{supAbs}\ (ZXY+,-) = \sum_{e_i \in C \cap (X)_{e_j \in \Omega}} (z_2 - z_i)(y_i - y_{\min})$$

$$\text{supAbs}\ (ZXY-,-) = \sum_{e_i \in C \cap (X)_{e_j \in \Omega}} (z_2 - z_i)(y_{\max} - y_i)$$

Notre objectif est de trouver tous les groupes *G* (*qualitatifs et quantitatifs*) dont un des supports absolus s_0 = supAbs(*ZM,+/-*) (*définitions 1 à 4*) diffère significativement du support attendu c'est-à-dire du support obtenu dans le cas où il n'y a pas de corrélation entre le motif *M* et la zone d'intérêt de la variable cible *Z*. Nous appellerons groupe atypique, tout groupe dont un des supports absolus s'écarte significativement du support absolu attendu.

Soit *S* la variable aléatoire dont s_0 est un support absolu et soit α le risque de première espèce[1].

Définition 5. Un groupe *G* (*qualitatif ou quantitatif*) est **atypique** si l'un de ses supports absolus s_0 diffère significativement du support absolu attendu, c'est-à-dire lorsqu'une des deux conditions est vérifiée :

condition 1 *: $Pr(S \leq s_0) \geq 1 - \alpha$*
condition 2 *: $Pr(S \leq s_0) \leq \alpha$*

Suivant la condition vérifiée, nous parlerons de groupe atypique positif (*condition 1*) ou groupe atypique négatif (*condition 2*).

Nous allons maintenant exposer la mesure utilisée pour extraire ces groupes atypiques. Pour cela, nous avons utilisé une mesure existante : l'intensité d'inclination.

3 Mesure d'extraction des groupes atypiques

Dans un premier temps, nous exposons la mesure existante, l'intensité d'inclination (Guillaume, 2002), et ensuite, nous allons voir comment nous l'avons utilisée afin d'extraire les groupes atypiques.

3.1 Intensité d'inclination

Soient deux variables quantitatives *Y* et *Z* prenant leurs valeurs dans respectivement les intervalles [y_{min}, y_{max}] et [z_{min}, z_{max}]. Soient y_i et z_i les valeurs prises par l'individu e_i pour respectivement les variables *Y* et *Z*. Soit *N* le nombre d'individus dans l'ensemble d'apprentissage Ω. Soient μ_Y et μ_Z les moyennes respectives des variables *Y* et *Z* et v_Y et v_Z les variances respectives des variables *Y* et *Z*.

[1] Le risque de première espèce α correspond au risque de rejeter à tort l'hypothèse d'absence de lien alors que celle-ci est vraie.

L'intensité d'inclination mesure si la valeur observée t_0 définie ci-dessous est significativement faible comparativement à la valeur attendue dans le cas où les deux variables *Y* et *Z* sont indépendantes.

$t_o = \sum_{i=1}^{N} (z_i - z_{\min})(y_{\max} - y_i)$

La variable aléatoire *T*, dont t_0 est une valeur observée, suit asymptotiquement la loi normale $\mathcal{N}(\mu, \sigma)$ avec

$\mu = N(\mu_Z - z_{min})(y_{max} - \mu_Y)$ et $\sigma^2 = N[v_Z v_Y + v_Y(\mu_Z - z_{min})^2 + v_Z(y_{max} - \mu_Y)^2]$.

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.

Afin de mesurer la *"petitesse"* de cet écart de façon croissante, l'indice $\varphi(Z,Y) = Pr(T > t_o)$ est retenu. Ainsi, la faiblesse de cet écart 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 : $\varphi(Z,Y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{t_0}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$

et la faiblesse de la valeur t_0 est significative si $\varphi(Z,Y) \geq 1-\alpha$.

3.2 Détection des groupes atypiques

L'intensité d'inclination mesure la *"petitesse"* de la valeur t_0 comparativement à ce que l'on obtiendrait si les deux variables étaient indépendantes. Nous avons donc une application directe de cet indice pour détecter si le support *supAbs(ZXY-,+)* est significativement faible en considérant comme ensemble d'apprentissage, non plus l'ensemble Ω, mais l'ensemble $C \cap (X)_{e_i \in \Omega}$. La faiblesse du support sera significative si $\varphi(Z,XY) \geq (1-\alpha)$.

Afin de mesurer si ce support absolu *supAbs(ZXY-,+)* est maintenant significativement élevé, il suffit de vérifier que le complément à *1* de l'intensité d'inclination est admissible au niveau de confiance $(1-\alpha)$. Un support élevé *supAbs(ZXY-,+)* sera significatif si $1-\varphi(Z,XY) \geq (1-\alpha)$.

Le *tableau 1* résume l'utilisation de l'intensité d'inclination pour détecter les groupes atypiques positifs et négatifs.

		Groupe atypique négatif	**Groupe atypique positif**
Groupe quantitatif	supAbs(*ZXY-,+*)	$\varphi(Z,XY) \geq (1-\alpha)$	$1-\varphi(Z,XY) \geq (1-\alpha)$
	supAbs(*ZXY+,+*)	$\varphi(Z,X(y_{max}+y_{min}-Y)) \geq (1-\alpha)$	$1-\varphi(Z,X(y_{max}+y_{min}-Y)) \geq (1-\alpha)$
	supAbs(*ZXY-,-*)	$\varphi(X(y_{max}+y_{min}-Y),Z) \geq (1-\alpha)$	$1-\varphi(X(y_{max}+y_{min}-Y),Z) \geq (1-\alpha)$
	supAbs(*ZXY+,-*)	$\varphi(XY,Z) \geq (1-\alpha)$	$1-\varphi(XY,Z) \geq (1-\alpha)$
Groupe qualitatif	supAbs(*ZX,-*)	$\varphi(X,Z) \geq (1-\alpha)$	$1-\varphi(X,Z) \geq (1-\alpha)$
	supAbs(*ZX,+*)	$\varphi(Z,1-X) \geq (1-\alpha)$	$1-\varphi(Z,1-X) \geq (1-\alpha)$

TAB. 1 – *Utilisation de l'intensité d'inclination pour détecter les groupes atypiques.*

4 Groupes atypiques intéressants

Après avoir défini la notion de groupes atypiques, et comment faire pour les extraire grâce à l'utilisation de l'intensité d'inclination, nous allons définir deux critères qui vont nous permettre d'obtenir des groupes atypiques intéressants.

Le premier critère va permettre de retenir des groupes atypiques dignes d'intérêt (*groupes atypiques fréquents*) et le second va s'assurer que chaque groupe atypique extrait est porteur d'une nouvelle information (*groupes atypiques informatifs*).

Groupes atypiques fréquents.

Le premier critère, *critère 1*, va s'assurer que le groupe atypique est digne d'intérêt, c'est-à-dire qu'il est vérifié par un certain nombre d'individus. Par analogie avec les motifs fréquents (Agrawal et al., 1996), nous appellerons ces groupes, des groupes atypiques fréquents.

Critère 1. Soit $seuil_1$ le support minimum défini par l'utilisateur. Un groupe atypique G est dit **fréquent** si le motif catégoriel X associé au groupe G est fréquent, autrement dit si $\Pr(X/Z) \geq seuil_1$.

Groupes atypiques informatifs.

Le deuxième critère, *critère 2*, va s'assurer que chaque groupe atypique extrait est informatif c'est-à-dire qu'il est porteur d'une nouvelle information, autrement dit que la même information n'est pas déjà présente dans un groupe atypique plus général.

Afin de définir ce critère, nous allons préciser ce que nous entendons par groupe plus général ou super-groupe.

Définition 6. Soient M et M' les motifs associés respectivement aux groupes G et G'. Le groupe G est un super-groupe du groupe G' si $M \subset M'$ c'est-à-dire si le motif M est inclus dans le motif M'. Nous dirons également que le groupe G' est un sous-groupe de G.

Ainsi, par exemple, le motif M=(*Sexe=″féminin″*) est inclus dans le motif M'=(*Sexe=″féminin″*)$\wedge$(*Profession=″cadre″*).

Dans la suite de cet article, tout nom de groupe suivi d'une apostrophe indiquera que nous sommes en présence d'un sous-groupe, c'est-à-dire d'un groupe où le motif qui lui est associé est composé d'au moins deux variables.

L'obtention de groupes atypiques positifs (*ou respectivement négatifs*) G' non intéressants intervient lorsqu'un de ses super-groupes G est très fortement présent (*ou respectivement très peu présent*) dans la zone étudiée de la variable cible, voire presque exclusivement présent (*ou respectivement absent*) dans cette zone. C'est le cas par exemple des personnes exerçant une profession libérale, qui ont fait de nombreuses années d'étude et qui, par conséquent, sont présentes uniquement dans la zone des très fortes valeurs de la variable *″nombre d'années d'étude″*. L'ajout de tout autre motif au motif profession=*″profession libérale″*, comme par exemple le motif sexe=*″féminin″*, restituera la même information puisque ce nouveau groupe positif G' (*femmes exerçant une profession libérale*) est un sous-ensemble des individus exerçant une profession libérale. La condition d'obtention de tels groupes atypiques informatifs peut se formaliser de la façon suivante :

Définition 7. Soit $seuil_2$ un seuil maximum défini par l'utilisateur.

Le groupe atypique positif (*ou respectivement négatif*) G' est **informatif** par rapport à une zone z de la variable cible Z si aucun de ses super-groupes positifs (*ou respectivement*

négatifs) *G* n'est présent de façon exclusive (*ou respectivement fortement absent*) dans cette même zone, c'est-à-dire n'a de support supérieur (*ou respectivement inférieur*) à un certain seuil :

Groupe positif : $\forall M \subset M'$ supAbs(ZM, z) $\leq seuil_2$

Groupe négatif : $\forall M \subset M'$ supAbs(ZM, z) $\geq seuil_2$

Selon la zone étudiée, nous avons deux supports possibles : supAbs(ZM,-) pour la zone des faibles valeurs de l'intervalle cible (*i.e.* z = -) et supAbs(ZM,+) pour la zone des fortes valeurs (*i.e.* z = +). Dans le cas où le motif *M* est un motif quantitatif et contient donc une variable quantitative, nous sommes en présence de deux autres supports : le support supAbs(ZM+, z) des fortes valeurs pour *Y* et le support supAbs(ZM-, z) des faibles valeurs pour *Y*.

Après avoir défini ces deux critères qui vont nous permettre d'obtenir des groupes intéressants, nous allons voir comment nous les avons utilisés afin de réduire l'espace de recherche des motifs.

5 Algorithme

Dans cette section, nous présentons l'algorithme dans le cas de la détection des groupes atypiques intéressants positifs pour la zone des fortes valeurs de l'intervalle cible. Nous pouvons transposer sans difficulté cette recherche au cas des groupes négatifs pour toute zone de l'intervalle ainsi que pour la zone des faibles valeurs dans le cas de groupes positifs.

L'algorithme d'extraction des groupes atypiques positifs présenté en *figure 1* est basé sur *Apriori* (Agrawal et al., 1996) et plus particulièrement sur la première partie de cet algorithme, à savoir l'extraction des motifs fréquents. Notre algorithme prend en entrée une table de données dans laquelle les variables qualitatives ont subi un codage disjonctif complet, une variable quantitative cible *Z* ainsi que la zone de l'intervalle d'étude [z_1, z_2], le risque de première espèce, un support minimum et un support maximum, et retourne l'ensemble des groupes atypiques positifs intéressants pour la zone concernée.

Comme l'algorithme *Apriori*, notre algorithme effectue un parcours par niveau du treillis de l'ensemble des parties de l'ensemble des motifs de *taille 1* (*i.e. motifs composés d'une seule variable*) et utilise l'anti-monotonie du support ainsi que la *propriété 1* présentée ci-dessous afin d'effectuer des coupures. La complexité de notre algorithme est identique à celle de l'algorithme *Apriori*, c'est-à-dire linéaire en nombre de passes sur la table de données mais exponentielle par rapport au nombre de motifs de *taille 1*.

Cette *propriété 1* utilise le *critère 2* présenté dans la *section 4*, qui nous indique que si un motif *M* est très fréquent (*ou respectivement très peu fréquent*) pour une zone *z* donnée alors tout sur-motif *M'* conduira à l'obtention d'un groupe *G'* non informatif.

Propriété 1.

Groupe positif : **si** supAbs(ZM, z) $\geq seuil_2$ **alors** $\forall M'\ M \subset M'\ G'(M')$ sera non informatif

Groupe négatif : **si** supAbs(ZM, z) $\leq seuil_2$ **alors** $\forall M'\ M \subset M'\ G'(M')$ sera non informatif

L'algorithme commence par rechercher les groupes atypiques positifs les plus généraux c'est-à-dire les groupes dont le motif associé est composé d'une seule variable (*étapes 1 à 3*). L'*étape 1* recherche les motifs qualitatifs fréquents. L'*étape 2* recherche les groupes atypiques positifs parmi les motifs qualitatifs fréquents et les variables quantitatives. L'*étape 3*

détecte les motifs fortement fréquents (*critère 2 de la section 4*) qui seront ensuite éliminés de l'ensemble des motifs fréquents servant à générer les motifs candidats.

Algorithme

Entrée : une table de données, la variable cible quantitative Z ainsi que la zone de l'intervalle $[z_1, z_2]$, le risque de première espèce α, un support minimum $seuil_1$ et un support maximum $seuil_2$.
Sortie : l'ensemble des groupes atypiques positifs intéressants donc l'ensemble des motifs associés aux groupes.

DEBUT
// **(1) Calcul des motifs qualitatifs X_i fréquents de *taille 1***
$LC_1=\{X_i$ / support$(X_i) \geq seuil_1$ et $|X_i|=1\}$
// **(2) Calcul des motifs *M* dont le groupe associé est atypique positif**
$LAC_1=\{X_i / X_i \in LC_1$ et $1\text{-}\varphi(Z, 1\text{-}X_i,) \geq (1\text{-}\alpha)\}$
$LAQ_1=\{Y_i(+/-) / 1\text{-}\varphi(Z, Y_i) \geq (1\text{-}\alpha)$ ou $1\text{-}\varphi(Z,_i(y_{max}+y_{min}\text{-}Y_i)) \geq (1\text{-}\alpha)\}$
// **(3) Détection des motifs X_i et $Y_i(+/-)$ fortement présents**
$LPC_1=\{X_i / X_i \in LAC_1$et supAbs$(ZX_i,+) \geq seuil_2\}$
$LPQ_1=\{Y_i(+/-) / Y_i(+/-) \in LAQ_1$ et supAbs$(ZY_i(+/-),+) \geq seuil_2\}$
$k=2$
Tant que $LC_{k-1} <> \varnothing$ **faire**
// **(4) Génération des motifs candidats de *taille k***
$CC_k=\{X_i / |X_i|=k\}$ à partir des $LC_{k-1} \setminus LPC_{k-1}$
$LQ_k=\{X_iY_i(+/-) / |X_i| = k\text{-}1$ et $|Y_i(+/-)|=1\}$
à partir des $LC_{k-1} \setminus LPC_{k-1}$ et des $LAC_1 \setminus LPQ_1$
// **(5) Calcul des motifs qualitatifs fréquents de *taille k***
$LC_k=\{X_i / X_i \in CC_k$ et support$(X_i) \geq seuil_1\}$
// **(6) Calcul des motifs *M* dont le groupe associé est atypique positif**
$LAC_k=\{X_i / X_i \in LC_k$ et $1\text{-}\varphi(Z, 1\text{-}X_i,) \geq (1\text{-}\alpha)\}$
$LAQ_k=\{X_iY_i(+/-) / X_iY_i(+/-) \in LQ_k$ et $1\text{-}\varphi(Z, X_iY_i) \geq (1\text{-}\alpha)$
ou $1\text{-}\varphi(Z, X_i(y_{max}+y_{min}\text{-}Y_i)) \geq (1\text{-}\alpha)\}$
// **(7) Détection des motifs fortement présents**
$LPC_k=\{X_i / X_i \in LAC_k$ et supAbs$(ZX_i,+) \geq seuil_2\}$
$k=k+1$
Fin Tantque
Retourner $\bigcup_{i=1..k} (LAC_i \cup LAQ_i)$
FIN

FIG. 1 – *Algorithme d'extraction des groupes atypiques positifs intéressants pour la zone des fortes valeurs de l'intervalle de la variable cible quantitative.*

Ensuite, les *étapes 4* à *7* vont extraire les sous-groupes atypiques positifs intéressants et ces quatre étapes vont être répétées pour chaque *niveau k* (*un niveau k correspond à la recherche des motifs composés de k variables*) jusqu'à ce que l'ensemble des motifs qualitatifs fréquents de niveau inférieur soit non vide (*car nous ne pourrions pas générer de motifs candidats*). L'*étape 4* génère à la fois : (1) des motifs qualitatifs candidats (*suivant le même principe que l'algorithme Apriori*) à partir des motifs fréquents de niveau inférieur privé des motifs fortement présents, et (2) des motifs quantitatifs candidats composés d'une variable

quantitative non fortement présente et d'un motif qualitatif fréquent de niveau inférieur non fortement présent également. L'*étape 5* calcule les motifs qualitatifs fréquents à partir des motifs candidats trouvés à l'*étape 4*. L'*étape 6* extrait les groupes atypiques positifs intéressants à partir des motifs fréquents. L'*étape 7* détecte les motifs fortement présents parmi les motifs qualitatifs afin de les supprimer de l'ensemble des motifs qualitatifs fréquents qui vont servir à générer les candidats de niveau supérieur.

L'algorithme a été implémenté en Java et intégré au logiciel libre d'extraction de connaissances WEKA (***W**aikato **E**nvironment for **K**nowledge **A**nalysis*) (Witten et Franck, 2005), logiciel développé par l'université de Waikato en Nouvelle-Zélande.

6 Expérimentations

Pour évaluer notre approche, la recherche des groupes atypiques a été effectuée sur la base de données IPUMS (*Integrated Public Use Microdata Series*), données du bureau de recensement américain. Ces données sont disponibles sur UCI KDD archive.

Cette base de données se compose de *88 443* individus décrits par *61* variables dont *29* sont des variables quantitatives. Après avoir effectué un codage disjonctif complet des *32* variables qualitatives, nous sommes en présence de *772* variables. Notre étude va se focaliser sur la variable quantitative cible : *"salaire"*. Cette variable prend ses valeurs dans l'intervalle [*0, 999 999*]. Cependant *23,98*% des individus possèdent la valeur *999 999*, ce qui ne modélise pas la réalité et correspond à une valeur non applicable pour l'unité statistique étudiée. C'est pourquoi nous avons décidé d'effectuer l'extraction sur l'intervalle [*0, 195 516*] (*la valeur 195 516 étant la valeur immédiatement inférieure à 999 999*). Les paramètres retenus pour cette extraction sont les suivants : risque de première espèce égal à *0,05* et support minimum positionné à *0,01*.

Voici quelques-uns des groupes atypiques découverts dans la zone des forts salaires :

Groupes atypiques qualitatifs

- Les hommes sont fortement présents dans la zone des forts salaires contrairement aux femmes qui sont moins représentées : G_1(*sexe= "masculin",*+)$^+$ et G_2(*sexe= "féminin",*+)$^-$.
- Cependant, pour certains métiers, comme par exemple les métiers codés *683* et *690*, les hommes ne sont plus fortement présents dans la zone des forts salaires mais au contraire sous-représentés : G_4(*sexe= "masculin",métier=683,*+)$^-$, G_5(*sexe= "masculin",métier=690,*+)$^-$.
- les individus naissant dans le lieu *042*, sont peu présents dans la zone des forts salaires contrairement aux individus qui naissent dans le lieu *006* : G_6(*lieuNaissance=042,*+)$^-$ et G_7(*lieuNaissance=006,*+)$^+$.
- les individus utilisant le moyen de transport codé *00* pour aller au travail, sont fortement présents dans la zone des forts salaires, contrairement à ceux qui utilisent le moyen codé *10* : G_8(*moyenTravail=00,*+)$^+$ et G_9(*moyenTravail=10,*+)$^-$.

Groupes atypiques quantitatifs

- Plus le statut de pauvreté est élevé, moins il y a d'individus dans la zone des forts salaires et plus le statut de pauvreté est faible, plus il y a d'individus dans cette zone : G_{10}(*statusPauvreté+,*+)$^-$ et G_{11}(*statusPauvreté-,*+)$^+$.
- Cependant, cette tendance s'inverse pour les individus qui exercent le métier codé *320* : G_{12}(*métier=320,statusPauvreté+,*+)$^+$ et G_{13}(*métier=320,statusPauvreté-,*+)$^-$.

- Plus le nombre de couples mariés est important dans le foyer familial, moins il y a d'individus dans la zone des forts salaires et plus le nombre de couples mariés est faible, plus il y a d'individus dans la zone des forts salaires : G_{14}(*nombreCouplesMariés*+,+)$^-$ et G_{15}(*nombreCouplesMariés*-,+)$^+$.
- La tendance précédente n'est plus vérifiée pour les foyers où le père de référence est un père biologique et un employé de *classe 2* .
G_{16}(*pèreAdoptif=0,classeEmployé=2,nombreCouplesMariés*+,+)$^+$.
- D'autres groupes plus complexes d'interprétation ont été mis en évidence :
G_{17}(*typeLogement=0,pèreAdoptif=0,forceTravail=2,dernièreAnnéeTravaillée=00, travailAnnéeDernière=2,nombreMèreDansLogement*+,+)$^-$
G_{18}(*lienMèreEnfant=0,lienPèreEnfant=0,scolarisé=1,nombreFamillesDansLogement*+,+)$^+$

7 Conclusion et perspectives

Dans cet article, nous avons proposé une technique d'extraction de groupes atypiques intéressants pour une variable quantitative cible, et plus particulièrement pour les faibles et fortes valeurs d'un intervalle de cette variable. Cette technique nous permet d'extraire des associations avec une variable quantitative en nous dispensant de l'étape de discrétisation des variables quantitatives. Cela élimine les erreurs liées à une discrétisation *a priori* et nous permet d'avoir une vue globale de l'association avec ces variables quantitatives et non plus un émiettement de la connaissance dû à cette transformation. La complexité du problème s'en trouve légèrement réduite, surtout pour les bases possédant de nombreuses variables quantitatives, puisqu'il n'y a pas multiplicité des variables. Cette extraction des groupes atypiques prend comme référence l'ensemble d'apprentissage. Il pourrait être intéressant d'introduire un autre paramètre : un ensemble de référence, ce qui nous permettrait de rechercher tous les groupes ayant un comportement identique ou différent de cet ensemble de référence.

Références

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.

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

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

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

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

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

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.

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

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.

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.

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

Summary

An interesting task in data analysis is the understanding of unexpected or atypical behaviors in a group of individuals. Which categories of individuals earn the higher salaries or, on the contrary, which ones earn the lower salaries? We present the problem of how data concerning atypical groups can be mined compared with a target quantitative attribute, like for instance the attribute "salary", and in particular for the high and low values of a user-defined interval. Our search therefore focuses on conjunctions of attributes whose distribution differs significantly from the learning set for the interval's high and low values of the target attribute. Such atypical groups can be found by adapting an existing measure, the intensity of inclination. This measure frees us from the transformation step of quantitative attributes, that is to say the step of discretization followed by a complete disjunctive coding. Thus, we propose an algorithm for mining such groups using pruning rules in order to reduce the complexity of the problem. This algorithm has been developed and integrated into the WEKA software for knowledge extraction. Finally we give an example of data extraction from the American census database IPUMS.