Des Règles d'Association Ordinales aux Règles d'Association "Classiques"

Sylvie Guillaume

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

Résumé. L'intensité d'inclination est une mesure qui permet d′extraire des règles directement sur les données ordinales sans avoir à les transformer par un codage disjonctif complet précédé d′une discrétisation pour les variables quantitatives. Ce nouveau type de règles, les règles d′association ordinales, dégagent les comportements généraux de la population et il est essentiel d'aller au plus près des individus en découvrant des règles spécifiques c'est-à-dire des règles vérifiées par des sous-ensembles d'individus. Cet article présente donc la technique pour extraire des règles d'association en partant des règles d'association ordinales, se libérant ainsi de l'étape de transformation des données et surtout en obtenant une discrétisation des variables quantitatives fonction du contexte c'est-à-dire fonction des variables auxquelles elles sont associées. L'étude se termine par une évaluation sur une base de données bancaires.

1 Introduction

Les règles d'association (Agrawal et al. 1993) permettent de détecter des implications intéressantes entre les variables binaires d'une base de données. Cependant, les algorithmes d'extraction (Mannila et al. 1994 ; Agrawal et al. 1996) de ces règles reposent sur des fréquences (*taux de couverture et probabilité conditionnelle*) et par conséquent, sont limités pour l'étude de variables ordinales (*variables qualitatives ordinales et variables quantitatives*). Srikant et Agrawal (Srikant et al. 1996) présentent des techniques pour discrétiser automatiquement les variables quantitatives afin d'utiliser les algorithmes précédents. C'est également une préoccupation de Miller et Yang (Miller 1997), que de déterminer les intervalles des variables quantitatives. Aumann et Lindel (Auman et al. 1999) ainsi que G.I. Webb (Webb 2001) proposent un nouveau type de règles, les règles d'impact, qui détectent des interactions intéressantes entre des combinaisons de variables binaires en prémisse et une variable quantitative en conclusion. Fukuda et al. (Fukuda et al. 1996) ainsi que Yoda et al. (Yoda et al. 1997) proposent des règles composées de deux variables quantitatives en prémisse et d'une variable binaire en conclusion. Pour finir, Guillaume (Guillaume 2002) propose également un nouveau type de règles, les règles d'association ordinales, qui détectent des implications entre des combinaisons de tout type de variables. Ces règles reposent sur une mesure d'intérêt, l'intensité d'inclination, qui évalue la petitesse du nombre d'individus qui violent la règle. Cette mesure évite l′étape de transformation des données ordinales, c′est-à-dire l′étape de codage disjonctif complet précédée de l′étape de discrétisation pour les variables quantitatives, évitant ainsi l′obtention d′un nombre important

de règles, règles pour la plupart faiblement significatives ou redondantes. Ces règles d′association ordinales extraites par cette mesure sont vérifiées par l′ensemble des individus de la base. Afin d′affiner notre connaissance du comportement des individus, il est essentiel d′extraire des règles sur des sous-ensembles d′individus. Cet article présente donc la technique pour extraire des règles d'association ″classiques″ en partant des règles d'association ordinales, se libérant ainsi de l'étape de transformation des données ordinales et surtout en obtenant une discrétisation des variables quantitatives fonction du contexte c'est-à-dire fonction des variables auxquelles elles sont associées.

Ainsi, cet article s'organise de la façon suivante. Dans la *section 2* nous rappelons la définition de l′intensité d′inclination et nous allons développer le principe et les critères qui permettent à cette mesure de retenir une règle d′association ordinale afin de justifier la démarche retenue en *section 3* pour extraire les règles sur des sous-ensembles d′individus. La *section 3* définit la procédure pour extraire les règles d′association ″classiques″ en partant des règles d′association ordinales. Dans la *section 4* nous évaluons cette technique d′extraction et une conclusion générale résume l'ensemble des points abordés et quelques perspectives sont envisagées pour la suite de ce travail.

2 Intensité d'inclination

Dans cette section, nous rappelons tout d′abord la définition de l'intensité d'inclination et nous allons ensuite approfondir le principe de cette mesure afin de comprendre la démarche retenue pour l′extraction des règles spécifiques détaillée en *section 3*.

2.1 Définition

L′intensité d′inclination (Guillaume 2002), mesure évaluant l′implication entre des conjonctions de variables ordinales, est une généralisation de l'intensité de propension (Lagrange 1997) (*indice évaluant la liaison implicative entre deux variables quantitatives à valeurs dans l'intervalle [0..1]*) et de l′intensité d′implication (Gras 1979) (*mesure évaluant la liaison implicative entre des conjonctions de variables binaires*).

Soient X et Y deux conjonctions de respectivement p et q variables ordinales. Nous posons $X = X_1,..,X_p$ et $Y = Y_1,..,Y_q$, où X_1, .., X_p, Y_1, .., Y_q sont des variables ordinales à valeurs x_{1_i} , .., x_{p_i} , y_{1_i} , .., y_{q_i} ($i \in \{1..N\}$, *N étant le nombre d'individus de la base de données ou population* Ω) dans respectivement les intervalles $[x_{1_{\min}}..x_{1_{\max}}]$, .., $[x_{p_{\min}}..x_{p_{\max}}]$, $[y_{1_{\min}}..y_{1_{\max}}]$, .., $[y_{q_{\min}}..y_{q_{\max}}]$. Afin de prendre en compte les variables qualitatives ordinales, un codage approprié des valeurs de celles-ci dans l'ensemble des réels doit être effectué.

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. Ces individus qui ne vérifient pas fortement la règle sont appelés contre-exemples.

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, ou mesure brute de non inclination, est défini de la façon suivante :

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

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

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

μ_{X_j} , μ_{Y_k} les moyennes respectivement des variables X_j et Y_k et

σ_{X_j} , σ_{Y_k} 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 :

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

avec $\mathrm{cov}(X_i, X_{i'}) = \mu_{X_i X_{i'}} - \mu_{X_i} \mu_{X_{i'}}$.

Si la probabilité $Pr(T \leq t_o)$ d'avoir un nombre inférieur ou égal à t_o est élevée, on peut 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 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 :

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

2.2 Principe

Soient r et s le nombre de valeurs distinctes prises respectivement par les variables X et Y dans la base de données ou population Ω. Soient $x_{\max} = \sum_{j=1}^{p} x'_{j_{\max}}$ et $y_{\min} = \sum_{k=1}^{q} y'_{k_{\min}}$ les valeurs respectivement maximale et minimale prises par les variables X et Y dans la population Ω.

Nous supposons que les différentes valeurs prises par X et Y sont ordonnées de la façon suivante : $x^{(1)} = x_{min} < x^{(2)} < \ldots < x^{(r-1)} < x^{(r)} = x_{max}$ et

$y^{(1)} = y_{min} < y^{(2)} < \ldots < y^{(s-1)} < y^{(s)} = y_{max}$.

Comme l'intensité d'inclination évalue la *"petitesse"* du nombre de contre-exemples, une comparaison entre les effectifs observés[1] n_{ij} et les effectifs théoriques $\frac{n_{i.}n_{.j}}{N}$ est effectuée ($n_{i.}$ *est la distribution marginale*[2] *de* $x^{(i)}$ *et* $n_{.j}$ *celle de* $y^{(j)}$).

Le *tableau 1* donne la contingence des écarts entre les effectifs observés et les effectifs théoriques pour les variables *X* et *Y* dans la base de données selon les *r* et *s* valeurs distinctes prises par ces deux variables.

	$y_{min}=y^{(1)}$	…	$y^{(j)}$	…	$y_{max}=y^{(s)}$	
$x_{min}=x^{(1)}$	$n_{11}-n_{.1}n_{1.}/N$	…	$n_{1j}-n_{.j}n_{1.}/N$	…	$n_{1s}-n_{.s}n_{1.}/N$	*0*
…	…	…	…	…	…	…
$x^{(i)}$	$n_{i1}-n_{.1}n_{i.}/N$	…	$n_{ij}-n_{.j}n_{i.}/N$	…	$n_{is}-n_{.s}n_{i.}/N$	*0*
…	…	…	…	…	…	…
$x_{max}=x^{(r)}$	$n_{r1}-n_{.1}n_{r.}/N$	…	$n_{rj}-n_{.j}n_{r.}/N$	…	$n_{rs}-n_{.s}n_{r.}/N$	*0*
	0	…	*0*	…	*0*	*0*

TAB 1 – *Tableau de contingence des écarts des variables X et Y.*

Remarque

La somme de toutes les lignes et de toutes les colonnes du tableau de contingence des écarts est nulle :

$$\forall i \in \{1,..,r\} \quad n_{i1} - \frac{n_{.1}n_{i.}}{N} + \ldots + n_{ij} - \frac{n_{.j}n_{i.}}{N} + \ldots + n_{is} - \frac{n_{.s}n_{i.}}{N} = n_{i.} - \frac{n_{i.}}{N}(n_{.1} + \ldots + n_{.j} + \ldots + n_{.s}) = 0$$

$$\forall j \in \{1,..,s\} \quad n_{1j} - \frac{n_{.j}n_{1.}}{N} + \ldots + n_{ij} - \frac{n_{.j}n_{i.}}{N} + \ldots + n_{rj} - \frac{n_{.j}n_{r.}}{N} = n_{.j} - \frac{n_{.j}}{N}(n_{1.} + \ldots + n_{i.} + \ldots + n_{r.}) = 0$$

L'intensité d'inclination garantit que les effectifs observés s'écartent significativement des effectifs théoriques et particulièrement dans la partie inférieure gauche du *tableau 1* et cet écart peut aller en s'atténuant au fur et à mesure que l'on s'éloigne de ce coin (*voir l'intensité du grisée du tableau 1 qui symbolise la force des écarts que l'on peut trouver*).

Exemple

Afin d'illustrer nos propos, nous prenons un exemple réel issu du domaine bancaire dont la base de données est présentée en *section 4*. La règle d'association ordinale *"*X_1*=action,* X_2*=disponible* $\rightarrow$ Y_1*=épargne logement"* a été extraite avec une intensité d'inclination égale à *0,9563*. Les variables X_1, X_2 et Y_1 représentent respectivement le nombre de comptes *"action"*, le nombre de compte *"disponible permanent"* et le nombre de *"prêt épargne logement"* ouvert par un foyer.

---

[1] L'effectif observé n_{ij} est le nombre d'individus vérifiant la valeur $x^{(i)}$ de la variable *X* ($X=x^{(i)}$) et la valeur $y^{(j)}$ de la variable *Y* ($Y=y^{(j)}$).

[2] La distribution marginale $n_{i.}=\sum_{k=1}^{s} n_{ik}$ de $x^{(i)}$ correspond à la distribution de $X=x^{(i)}$ sans tenir compte de *Y* et la distribution marginale $n_{.j}=\sum_{k=1}^{r} n_{kj}$ de $y^{(j)}$ correspond à la distribution de $Y=y^{(j)}$ sans tenir compte de *X*.

Le *tableau 2* représente le tableau de contingence des variables X ($X=X_1,X_2$) et Y ($Y=Y_1$).

	Y=0	***Y=1***	***Y=2***	***Y=3***	***Y=4***	***Y=5***	***Y=6***	***Total***
X=0	33 473	1 565	538	89	26	3	0	*35694*
X=1	8 563	813	277	55	16	1	2	*9727*
X=2	1 218	184	57	23	6	1	0	*1489*
X=3	149	22	12	4	2	0	0	*189*
X=4	7	2	1	0	0	0	0	*10*
X=5	1	1	1	0	0	0	0	*3*
Total	*43411*	*2587*	*886*	*171*	*50*	*5*	*2*	*47112*

TAB 2 – *Tableau de contingence pour l'exemple bancaire.*

Les variables X_1 et X_2 prennent respectivement leurs valeurs dans les ensembles {*0,1,2,3,4,5*} et {*0,1,2*} et la variable Y dans l'ensemble {*0,1,2,3,4,5,6*}. Il n'y a pas d'individu qui possède les trois combinaisons suivantes : (*4* comptes action, *2* comptes disponible), (*5* action, *1* disponible) et (*5* action, *2* disponible), ce qui explique pourquoi on ne trouve pas les valeurs *6* et *7* pour la variable X.

Le *tableau 3* représente le tableau de contingence des écarts entre les effectifs observés et théoriques des variables X et Y de la base de données bancaires.

	Y=0	***Y=1***	***Y=2***	***Y=3***	***Y=4***	***Y=5***	***Y=6***	
X=0	583	-395	-133	-41	-12	-1	-2	*-1*
X=1	-400	279	94	20	6	0	2	*1*
X=2	-154	102	29	18	4	1	0	*0*
X=3	-25	12	8	3	2	0	0	*0*
X=4	-2	1	1	0	0	0	0	*0*
X=5	-2	1	1	0	0	0	0	*0*
	0	*0*	*0*	*0*	*0*	*0*	*0*	*0*

TAB 3 – *Tableau de contingence des écarts pour l'exemple bancaire.*

Ainsi, pour *X=5* et *Y=0* on a deux individus en moins par rapport à ce qui est attendu et au contraire pour *X=1* et *Y=1*, on a *279* individus en plus par rapport à ce qui est attendu sous l'hypothèse que X et Y soient indépendantes.

Nous vérifions qu'il y a statistiquement peu d'individus qui vérifient à la fois beaucoup de comptes *"action"* et de comptes *"disponible permanent"* et peu de *"prêts épargne logement"*.

3 Spécialisation des règles d'association ordinales

Cette section va définir la procédure pour extraire les liaisons significatives sur des sous-ensembles de la population. Nous partons des règles d'association ordinales extraites sur l'ensemble de la population pour aller vers des règles vérifiées par des catégories d'individus.

C'est une recherche descendante découpée en deux étapes, chaque étape allant vers un degré de spécialisation plus important. La première étape va nous permettre d'obtenir des règles d'association ordinales spécifiques et la seconde étape des règles d'association "classiques".

Nous nous limitons à l'extraction des règles d'association ordinales composées d'une seule variable en conclusion (*comme (Webb 2001)*), les règles ayant plus d'une variable en conclusion sont plus difficilement exploitables par l'utilisateur.

3.1 Première étape de spécialisation

La première étape va permettre d'obtenir des règles ordinales du type $X=[x^{(i1)}..x^{(i2)}] \rightarrow Y=[y^{(j1)}..y^{(j2)}]$ avec $(x^{(i1)}, x^{(i2)}) \in [x^{(1)}..x^{(r)}]^2$ et $(y^{(j1)}, y^{(j2)}) \in [y^{(1)}..y^{(s)}]^2$.

Principe

De par les propriétés du tableau de contingence des écarts (*somme de chaque ligne et de chaque colonne nulle*), nous sommes schématiquement en présence d'un tableau du type de celui qui est représenté par la *figure 1*, dans les cas les plus simples évidemment c'est-à-dire lorsque le nombre de valeurs distinctes pour respectivement X et Y est de l'ordre d'une dizaine de valeurs. Dans les cas les plus compliqués, nous retrouvons ce type de tableau mais avec un nombre plus important de zones d'écarts positifs et d'écarts négatifs.

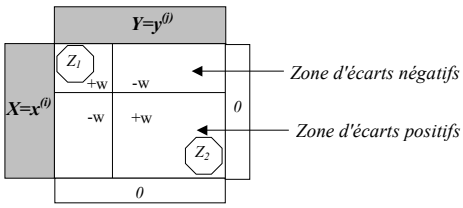

FIG. 1 - *Exemple de tableau de contingence des écarts pour les variables X et Y.*

Soient N_{ep} le nombre total des écarts positifs entre les effectifs observés et théoriques et m le nombre de zones Z_k ($k \in \{1,..,m\}$) d'écarts positifs. Pour le tableau représenté par la *figure 1*, $N_{ep}=2w$ et $m=2$ et pour le *tableau 3*, $N_{ep}=1167$ et $m=2$.

De par les expériences menées, le tableau des écarts possède de nombreuses cases ($X=x^{(i)}$, $Y=y^{(j)}$) dont la valeur est égale à zéro, c'est pourquoi nous allons rechercher dans toutes les zones Z_k, le plus grand rectangle R_k ne possédant aucune valeur nulle et couvert par un nombre minimum d'écarts c'est-à-dire que R_k doit vérifier un pourcentage de couverture P_{ep} supérieur à un seuil minimal min_{ep} fixé par l'utilisateur ($P_{ep}=n_{ek}/N_{ep}$ *avec* n_{ek} *le nombre d'écarts positifs du rectangle* R_k.) Nous définissons ce rectangle R_k par les deux points suivants : le point supérieur gauche P_g ($X=x^{(i1)}, Y=y^{(j1)}$) et le point inférieur droit P_d ($X=x^{(i2)}, Y=y^{(j2)}$). La connaissance de ces deux points va nous permettre ainsi de dégager la règle d'association ordinale spécifique $X=[x^{(i1)}..\ x^{(i2)}] \rightarrow Y=[y^{(j1)}..\ y^{(j2)}]$. Pour avoir une meilleure connaissance de cette règle, nous pouvons calculer son taux de couverture T_c qui est égal à $\frac{\left|\left(X=[x^{(i1)}..x^{(i2)}]\right)_{e \in \Omega} \cap \left(Y=[y^{(j1)}..y^{(j2)}]\right)_{e \in \Omega}\right|}{N}$.

La procédure d'extraction de ces règles est résumée par l'algorithme *RRAOS*.

Algorithme *RRAOS*

Entrée : Ensemble des règles d'association ordinales *RAO*
ROS=∅ *%initialisation de l'ensemble des règles d'association ordinales spécifiques*
Pour chaque règle d'association ordinale de *RAO*
- Calcul du tableau de contingence des écarts
- Recherche des *m* zones Z_k
- **Pour** chaque zone Z_k
  - Recherche des rectangles $R_k[P_g(X=x^{(i1)},Y=y^{(j1)}), P_d(X=x^{(i2)},Y=y^{(j2)})]$
  - **Pour** chaque rectangle R_k
    - **Si** $P_{ep}(R_k) \geq min_{ep}$
      $RO=[RO ; X=[x^{(i1)}..\ x^{(i2)}] \rightarrow Y=[y^{(j1)}..\ y^{(j2)}] \quad P_{ep}(R_k) \quad T_c]$
    - F**in Si** % $P_{ep}(R_k) \geq min_{ep}$
  - **Fin Pour** *%chaque rectangle* R_k
- **Fin Pour** *%chaque zone* Z_k

Fin Pour *%chaque règle d'association ordinale de RAO*
Sortie : Ensemble des règles d'association ordinales spécifiques *RAOS*

Exemple

Si nous reprenons l'exemple bancaire, le déroulement de l'algorithme *RRAOS* va extraire deux règles ordinales spécifiques qui sont résumées dans le *tableau 4*.

Entrée : *RO*={"X_1=*action*, X_2=*disponible* → Y_1=*épargne logement*"}			
Zones Z_k	***Rectangles*** R_k	$P_{ep}(R_k)$	T_c
Z_1	R_1 [P_g (X=0,Y=0),P_d (X=0,Y=0)]	0,50	0,71
Z_2	R_2 [P_g (X=1,Y=1),P_d (X=3,Y=4)]	0,49	0,03
Sortie : *ROS*={"X=0 → Y=0" P_{ep}=0,50 T_c=0,71 ; "X=[1..3] → Y=[1..4]" P_{ep}=0,49 T_c=0,03 }			

TAB 4 – *Déroulement de l'algorithme RRAOS pour l'exemple bancaire.*

Un traitement ultérieur (*utilisation de mesures subjectives*) éliminerait la première règle "$X=0 \rightarrow Y=0$" jugée inintéressante par l'utilisateur.

3.2 Deuxième étape de spécialisation

La seconde étape va nous permettre d'obtenir des règles d'association à partir des règles d'association ordinales spécifiques. Nous poursuivons notre recherche en allant au plus près des individus et des variables initiales X_i ($i \in \{1..p\}$).

Nous rappelons que nous nous limitons aux règles composées d'une variable en conclusion.

Soient r_1, .., r_p le nombre de valeurs distinctes prises respectivement par les variables X_1, .., X_p et s_1, .., s_q le nombre de valeurs distinctes prises respectivement par les variables Y_1, .., Y_q.

Cette étape va nous permettre d'obtenir des règles du type
$X_1=[x_1^{(i1)}..x_1^{(i2)}] \wedge .. \wedge X_p=[x_p^{(i1)}..x_p^{(i2)}] \rightarrow Y=[y^{(j1)}..y^{(j2)}]$ avec

$(x_1^{(i1)}, x_1^{(i2)}) \in [x_1^{(1)}..x_1^{(r1)}]^2, \ldots, (x_p^{(i1)}, x_p^{(i2)}) \in [x_p^{(1)}..x_p^{(rp)}]^2$.

Principe

Nous allons développer le tableau de contingence des variables X et Y pour faire apparaître les variables initiales X_i ($i \in \{1..p\}$).

Soit $c(i)$ le nombre de combinaisons pour la ligne $X=x^{(i)}$ ($i \in \{1..r\}$) du tableau de contingence des variables X et Y telle que la somme des p valeurs $X_1=x_1^{(i1)k}$, .., $X_p=x_p^{(ip)k}$ ($i_1 \in \{1..r_1\}$ ... ($i_p \in \{1..r_p\}$) et $k \in \{1..c(i)\}$) soit égale à $x^{(i)}$.

$$\forall k \in \{1..c(i)\} \quad \sum_{j=1}^{p} x_j^{(i_j)} = x^{(i)}$$

Le *tableau 5* représente ce développement de la ligne $X=x^{(i)}$.

	c(i) combinaisons			$Y=y^{(1)}$	…	$Y=y^{(j)}$	…	$Y=y^{(s)}$
$X=x^{(i)}$	$X_1=x_1^{(i1)1}$	..	$X_p=x_p^{(ip)1}$	n_{i11}	…	n_{ij1}	…	n_{is1}
	…	..	…	…	…	…	…	…
	$X_1=x_1^{(i1)c(i)}$	..	$X_p=x_p^{(ip)c(i))}$	$n_{i1c(i)}$	…	$n_{ijc(i)}$	…	$n_{isc(i)}$
	Total			n_{i1}	…	n_{ij}	…	n_{is}

TAB 5 – *Développement de la ligne $X=x^{(i)}$ du tableau de contingence.*

Le *tableau 6* développe le tableau de contingence de l'exemple bancaire (*c'est-à-dire le tableau 2*). Nous nous sommes limités aux valeurs qui nous intéressent à savoir les valeurs $X=1$, $X=2$ et $X=3$ puisqu'on étudie la règle d'association ordinale spécifique $X=[1..3] \rightarrow Y=[1..4]$.

		Y=0	***Y=1***	***Y=2***	***Y=3***	***Y=4***	***Y=5***	***Y=6***	***Total***
	X=0	33 473	1 565	538	89	26	3	0	*35694*
$X_1=0,X_2=1$		5 432	564	186	36	11	0	1	*6230*
$X_1=1,X_2=0$	***X=1***	3 131	249	91	19	5	1	1	*3497*
$X_1=0,X_2=2$		0	1	0	0	0	0	0	*1*
$X_1=1,X_2=1$	***X=2***	509	98	34	17	2	1	0	*661*
$X_1=2,X_2=0$		709	85	23	6	4	0	0	*827*
$X_1=1,X_2=2$		0	0	0	0	0	0	0	*0*
$X_1=2,X_2=1$	***X=3***	104	16	8	3	2	0	0	*133*
$X_1=3,X_2=0$		45	6	4	1	0	0	0	*56*
	X=4	7	2	1	0	0	0	0	*10*
	X=5	1	1	1	0	0	0	0	*3*
	Total	*43411*	*2587*	*886*	*171*	*50*	*5*	*2*	*47112*

TAB 6 – *Tableau de contingence partiellement développé pour l'exemple bancaire.*

Soit II_1 la valeur de l'intensité d'inclination pour la règle d'association ordinale spécifique $X=[x^{(i1)}..x^{(i2)}] \rightarrow Y=[y^{(j1)}..y^{(j2)}]$.

L'objectif est de déterminer pour la ligne $X=x^{(i)}$, la ou les combinaison(s) $X_1=x_1^{(i1)k}$, .., $X_p=x_p^{(ip)k}$ significatives dans l'apparition de la règle d'association ordinale spécifique. Pour

cela, nous allons éliminer chaque combinaison du tableau de contingence développé et calculer la nouvelle valeur de l'intensité d'inclination II_2. Si cette nouvelle valeur est suffisamment inférieure à II_1, cela signifie que cette combinaison est déterminante dans l'extraction de la règle. Afin de mesurer cet écart entre les deux valeurs de l'intensité d'inclination, nous allons calculer le rapport II_1/II_2. Ce rapport doit être supérieur à un seuil minimal min_r fixé par l'utilisateur (*seuil devant obligatoirement être supérieur à 1 pour que cette combinaison soit jugée déterminante*).

Nous pouvons affiner la connaissance de la règle d'association en calculant son taux de couverture T_c.

La procédure d'extraction de ces règles est résumée par l'algorithme *RRA*.

Algorithme *RRA*

Entrée : Ensemble des règles d'association ordinales spécifiques *RAOS*
$RA=\varnothing$ %*initialisation de l'ensemble des règles d'association RA*
Pour chaque règle de *RAOS* d'intensité d'inclination II_1
- Calcul du tableau de contingence étendu
- **Pour** chaque ligne $X=x^{(i)}$ concernée par la règle de *RAOS*
  - **Pour** chaque combinaison de la ligne $X=x^{(i)}$
    - Suppression de la combinaison du tableau de contingence
    - Calcul de l'intensité d'inclination II_2
    - **Si** $II_1/II_2 \geq min_r$
      RA=[RA ; $X_1=x_1^{(i1)}$, .., $X_p=x_p^{(ip)} \rightarrow Y=[y^{(j1)}..y^{(j2)}]$ T_c]
    Fin Si % $II_1/II_2 \geq min_r$
    F**in Pour** %*chaque combinaison de la ligne* $X=x^{(i)}$
  Fin Pour %*chaque ligne* $X=x^{(i)}$ *concernée par la règle de RAOS*
Fin Pour %*chaque règle d'association ordinale spécifique de RAOS*
Sortie : Ensemble des règles d'association *RA*

Exemple

Le *tableau 7* résume le déroulement de l'algorithme *RRA*. Nous fixons le seuil minimal min_r égal à *1,005*.

Combinaison	II_2	II_1/II_2	***Règle d'association***
$X_1=0,X_2=1$	0,9213	1,0380	$X_1=0 \wedge X_2=1 \rightarrow Y=[1..4]$
$X_1=1,X_2=0$	0,9530	1,0034	
$X_1=0,X_2=2$	0,9561	1,0002	
$X_1=1,X_2=1$	0,9218	1,0374	$X_1=1 \wedge X_2=1 \rightarrow Y=[1..4]$
$X_1=2,X_2=0$	0,9533	1,0032	
$X_1=1,X_2=2$	0,9563	1,0000	
$X_1=2,X_2=1$	0,9487	1,0080	$X_1=2 \wedge X_2=1 \rightarrow Y=[1..4]$
$X_1=3,X_2=0$	0,9547	1,0017	

TAB 7 – *Déroulement de l'algorithme RRA sur l'exemple bancaire.*

La règle d'association déduite de cet algorithme est donc $X_1=[0..2] \wedge X_2=1 \rightarrow Y=[1..4]$ avec un taux de couverture T_c égal à *0,1491*.

4 Evaluation sur des données bancaires

Dans cette section, nous présentons les règles d'association découvertes sur une base de données bancaire. Tout d'abord nous décrivons la base de données et ensuite nous donnons quelques règles d'association qui ont été découvertes.

La base de données bancaire se compose de *47 112* individus décrits par *44* variables quantitatives.

Les variables peuvent être répertoriées en trois catégories :

- Informations concernant le client (*âge, ancienneté, ...*),
- Informations sur les différents comptes du client (*actions, PEL, ...*),
- Statistiques sur les différents comptes (*montant des ressources, ...).*

La caractéristique de cette base de données est le faible pourcentage des individus qui possèdent un produit financier. L'exemple bancaire de la *section 3* le montre puisque *71,05*% des individus ne possèdent pas *X* et *Y*, *75,76*% ne possèdent pas *X* et *92,14*% ne possèdent pas *Y*. Par conséquent, l'extraction des règles d'association reposant sur le taux de couverture (*ou support*) et la probabilité conditionnelle (*ou confiance*) conduit à un nombre prohibitif de règles inintéressantes du type $"X=0 \rightarrow Y=0"$, $"X=1 \rightarrow Y=0"$, $"X=2 \rightarrow Y=0"$, $"Y=1 \rightarrow X=0"$, … Pour ce type de données, il est nécessaire de recourir à d'autres techniques.

708 règles d'association ont été extraites avec un niveau de confiance égal à *0,95* pour l'intensité d'inclination et un seuil minimal min_r égal à *1,005*. Nous nous sommes limités à l'extraction de règles composées au maximum de trois variables en prémisse.

Ainsi par exemple, nous avons détecté la règle d'association ordinale $"X_1=obligation, X_2=action \rightarrow Y_1=LER\text{-}PER\text{-}PEP\ Assurance"$ avec une intensité d'inclination égale à *0,9632*. Les variables X_1, X_2 et Y_1 représentent le nombre de comptes ouverts pour le produit financier mentionné et ces variables prennent leurs valeurs dans respectivement $[0..4] \cup \{6\}$, $[0..5]$ et $[0..11]$. La règle d'association ordinale spécifique $"X=[1..4] \rightarrow Y_1=[1..3]"$ a ensuite été extraite avec un taux de couverture T_c de *0,0381*. Cette règle spécifique a ensuite conduit à l'extraction de deux règles d'association $"X_1=0, X_2=[1,2] \rightarrow Y_1=[1..3]"$ avec $Tc=0,0189$ et $"X_1=1, X_2=[0..2] \rightarrow Y_1=[1..3]"$ avec $T_c=0,0184$. Nous vérifions que la variable *"action"* est présente dans ces règles avec des intervalles différents de l'exemple qui a servi à expliquer la technique proposée dans la *section 3*.

Autre exemple, la règle "$X_1=LER\text{-}PER\text{-}PEP\ Assurance, X_2=action, X_3=Credimatic \rightarrow Y_1=prêt\ épargne\ logement$" a été extraite avec une intensité d'inclination égale à *0,9605*. Ces variables prennent leurs valeurs dans respectivement $[0..11]$, $[0..5]$, $[0..6]$ et $[0..6]$. La règle d'association spécifique "$X=[1..5] \rightarrow Y_1=[1..3]$" avec $T_c=0,0371$ est ensuite dégagée et donne naissance aux règles d'association "$X_1=0, X_2=0, X_3=[1..2] \rightarrow Y_1=[1..3]$" avec $T_c=0,0127$ et $"X_1=1, X_2=1, X_3=0 \rightarrow Y_1=[1..3]"$ avec $T_c=0,0022$.

5 Conclusion

Les algorithmes d'extraction de règles d'association nécessitent une transformation de l'ensemble des variables initiales en des variables binaires. Cette étape a non seulement un coût élevé (*la complexité de ces algorithmes croît exponentiellement avec le nombre de variables*), délivre un nombre très important de règles faiblement expressives avec de fortes redondances mais aussi fige la découverte des règles qui font intervenir les variables quantitatives puisque les intervalles sont déterminés lors de la phase de discrétisation c'est-à-dire avant l'extraction.

Cet article propose une technique d'extraction de règles d'association qui se libère de l'étape de transformation des variables et qui permet une discrétisation des variables quantitatives pendant la phase d'extraction. La discrétisation obtenue est donc fonction du contexte c'est-à-dire fonction des autres variables intervenant dans la règle. La technique proposée est particulièrement adaptée aux données peu denses puisque l'extraction des règles ne repose pas sur le support mais sur l'importance de l'écart par rapport à l'indépendance.

Nous allons poursuivre ce travail par la proposition d'une nouvelle mesure ordinale, *le coefficient d'inclination*, afin de mieux prendre en compte la spécificité des variables qualitatives ordinales. Une comparaison avec les autres travaux doit également être menée.

Références

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

Agrawal R., Mannila H., Srikant R., Toivonen H. and 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. p. 307-328, 1996.

Auman Y., Lindell Y. (1999), A Statistical Theory for Quantitative Association Rules, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 99) p. 261-270, 1999.

Fukuda T., Morimoto Y., Morishita S., and Tokuyama T. (1997), Data mining using two-dimensional optimized association rules : Scheme, algorithms, and visualization, Proceedings of ACM SIGMOD International Conference Management of Data, p. 452-461, Tucson, AZ, 1997.

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, Octobre 1979.

Guillaume S. (2002), Découverte de Règles d'Association Ordinales, Actes des Journées Francophones d'Extraction et de Gestion des Connaissances (EGC'2002), 19-23 janvier 2002, pp. 29-40, Hermès Science Publications, Paris, ISBN 2-7462-0406-1.

Lagrange J.B. (1997), Analyse Implicative d'un Ensemble de Variables Numériques, Application au Traitement d'un Questionnaire à Réponses Modales Ordonnées, rapport interne Institut de Recherche Mathématique de Rennes, Prépublication 97-32 Implication Statistique, Décembre1997.

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, p. 181-192, Seattle, Washington, 1994.

Miller R.J., Yang Y. (1997), Association rules over interval data, Proceedings of ACM SIGMOD International Conference Management of Data, p. 452-461, Tucson, AZ, 1997.

Srikant R., Agrawal R. (1996), Mining quantitative association rules in large relational tables, Proceedings 1996 ACM-SIGMOD International Conference Management of Data, Montréal, Canada, June 1996.

Webb G.I. (2001), Discovering associations with numeric variables, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 383-388, San Francisco, CA, USA, August 26-29, 2001.

Yoda K., Fukuda T., Morimoto Y., and Tokuyama T. (1997), Computing optimized rectilinear regions for associations rules, Proceedings 3rd International Conference on Knowledge Discovery and Data Mining, p. 96-103, Newport Beach, California, 1997.

Summary

Intensity of inclination, a suitable way of measuring conjunctions of ordinal attributes, evaluates whether the number of transactions not strongly verifying the rule $X \rightarrow Y$ is significantly small compared to the expected number of transactions under the assumption that X and Y are independent. This objective rule-interest measure allows us to extract implications on databases without having to go through the step of transforming the initial set of attributes into binary attributes, thereby avoiding a prohibitive number of weakly significant rules with many redundancies. This new kind of rules, ordinal association rules, brings out the overall behavior of the population and this study has to be extended with the exploration of specific ordinal association rules in order to refine our analysis and to extract behaviors in sub-sets. This paper focuses on the mining of association rules based on extracted ordinal association rules in order to, on the one hand remove the discretization step of numeric attributes and the step of complete disjunctive coding, and on the other hand obtain a variable discretization of numeric attributes i.e. dependent on association of attributes. An evaluation of an application to some banking data ends up the study.