



Classification hiérarchique orientée en ASI

Pascale KUNTZ*

*Laboratoire d'Informatique de Nantes Atlantique
 Site Ecole Polytechnique de l'Université de Nantes
 La Chantrerie
 BP 50602
 44306 Nantes cedex 3
 France
 Pascale.Kuntz@polytech.univ-nantes.fr

Résumé. Cette communication est une synthèse de résultats nouveaux en ASI qui étendent la notion classique de quasi-implication (« quand a_i est présent alors généralement a_j est également présent ») au concept de R-règles (règles de règles) où la prémisse et la conclusion peuvent être elles-mêmes des quasi-implications. Nous définissons une nouvelle mesure pour valider la pertinence statistique de ces R-règles et développons le concept de « hiérarchie orientée » pour pouvoir les structurer. Afin de faciliter l'interprétation de cette nouvelle structure, nous proposons, suivant la méthodologie classique en classification hiérarchique, un critère de sélection des niveaux de la hiérarchie orientée les plus pertinents.

1 Introduction

A l'origine, l'analyse statistique implicative a été développée pour traiter des questions de relations que posait la didactique des mathématiques. Une des relations paradigmatiques rencontrée dans ce contexte était « si une question est plus complexe qu'une autre, alors tout élève qui réussit la première doit *a priori* réussir la seconde ». Inspiré par les travaux de I. Lerman sur l'analyse de la vraisemblance du lien (Lerman 1981), R. Gras a alors défini un indice à fondement probabiliste, l'intensité d'implication, pour quantifier l'étonnement de l'apparition d'une relation de quasi-implication de type « si on a a_i , alors on a presque sûrement a_j » (Gras 1979). La stricte implication de la logique classique est ici relaxée puisque l'on admet quelques contre-exemples dès lors que ceux-ci ne remettent pas en cause la tendance générale ; pour simplifier on parlera dans la suite de « règles ». L'indice d'intensité d'implication permet d'évaluer la pertinence des tendances implicatives associées à un tableau croisé classique *Individus* \times *Attributs*. Cependant, comme l'avait indiqué R. Gras dès le début de ses travaux (Gras et al 1996a), pour apprécier la qualité sémantique des règles, il convient de considérer simultanément l'ensemble de toutes les variables car « *c'est tout le réseau de relations qui permet de dégager une structure sémantique* ». D'où la nécessité d'une méthode multidimensionnelle qui puisse rendre compte du caractère intrinsèquement dissymétrique des faits analysés.

Pour représenter un ensemble S de règles sous une forme organisée, une approche souvent employée pour ses facilités de mise en œuvre et d'interprétation, consiste à modéliser les relations sur S par un graphe orienté. Dans la représentation la plus simple, les sommets du graphe sont les prémisses et les conclusions des règles, et les arcs représentent la tendance implicative (Rostam 1981). Un tel graphe permet, sans légende, d'appréhender les classes d'équivalence de la relation sur S « avoir une prémisse ou une conclusion commune ». En revanche, hormis la transitivité quand elle existe, il permet plus difficilement de déduire d'autres relations. Dans certains cas basés essentiellement sur le recours à la probabilité conditionnelle comme mesure de pertinence, cette limitation a conduit au développement d'autres modèles de graphes, permettant des inférences déductives sous forme de chemins (Horschka et Krögsen 1991, Lehn 2000).



Une autre voie consiste à rechercher une structuration d'un autre ordre basée sur des modèles de classification. Si l'on disposait d'un système implicatif complet sur l'ensemble A des attributs, l'arsenal mathématique des familles de Moore et des opérateurs de fermeture permettrait d'accéder, en particulier, à des modèles hiérarchiques (Domenach et Leclerc 2003). Cependant, dans le cadre de l'analyse statistique implicative, les ensembles de règles ne vérifient pas les propriétés requises.

Sans connaissance préalable d'un modèle spécifiquement approprié, une démarche classique utilisée notamment dans le contexte de l'Extraction de Connaissances dans les Données consiste à classifier S en recherchant un sous-ensemble de l'ensemble $P(S)$ des parties de S . La majorité des travaux publiés recherchent des partitions (Lent et al 1997, Toivonen et al 1995). Une dissimilarité est définie sur $S \times S$ ou sur le produit cartésien de l'ensemble des prémisses et celui des conclusions, et les partitions de S sont obtenus par des algorithmes de classification automatique.

Prolongeant un travail initié par R. Gras et A. Lahrer (Gras et Lahrer 1992), nous avons privilégié une autre voie qui permet non seulement de structurer certaines règles pertinentes, mais également de découvrir de nouvelles relations implicatives entre ces règles sous la forme $R \rightarrow R'$ où la prémisses R et la conclusion R' peuvent être elles-mêmes des règles. Le modèle proposé, appelé « hiérarchie orientée », est une extension du modèle hiérarchique classique sur l'ensemble des parties de A à un ensemble de règles : les niveaux de la hiérarchie orientée sont des règles ou des R-règles, et contrairement au modèle hiérarchique classique où l'ensemble A est dans la hiérarchie, une hiérarchie orientée ne contient que des règles significatives selon un critère statistique que nous avons défini.

Cette communication présente une synthèse de résultats récents obtenus en collaboration avec R. Gras (Gras et Kuntz 2005) et J.-C. Régner (Gras et al 2004) sur la construction des hiérarchies orientées et leur interprétation. Dans la première partie, nous présentons le concept de hiérarchie orientée dans un cadre théorique général. Dans la deuxième partie, nous introduisons dans le cadre de l'analyse statistique implicative le critère de validation des R-règles. L'algorithme de construction d'une hiérarchie orientée est décrit dans la troisième partie et quelques propriétés en sont déduites. Une discussion est ensuite proposée sur la significativité des niveaux d'un tel modèle.

2 Structuration des R-règles par une hiérarchie orientée

Dans la suite, nous considérons un ensemble I de n individus décrits par un ensemble $A = \{a_1, a_2, \dots, a_n\}$ de p attributs binaires.

A titre d'exemple, considérons un ensemble $A = \{a_1, a_2, a_3, a_4, a_5\}$ de 5 attributs. La figure 1 est une illustration d'une hiérarchie orientée H_A que l'on pourrait construire sur A . Les éléments de cette hiérarchie, appelés « classes » par analogie avec une hiérarchie classique, sont k -permutations de A , $k \leq 5$, qui satisfont des propriétés spécifiques d'emboîtement :

$$H_A = \{a_1, a_2, a_3, a_4, a_5, a_2a_3, a_5a_4, a_1a_5a_4\}$$

Chaque classe non élémentaire s'écrit comme une concaténation de deux classes de H_A : par exemple, $a_1a_5a_4$ est la « concaténation » des classes a_1 et a_5a_4 .

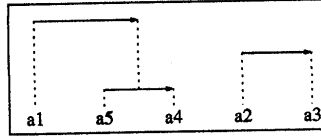


FIG. 1 – Représentation d'une hiérarchie orientée sur un ensemble $A = \{a_1, a_2, a_3, a_4, a_5\}$

Ces classes sont associées à des R-règles : la classe $a_1a_5a_4$ est associée à la R-règle $(a_1 \rightarrow (a_5 \rightarrow a_4))$. Cette association est univoque et dépend des conditions d'emboîtements sur les classes. Ainsi, la classe $a_1a_5a_4$ ne pourrait pas être associée à la R-règle $(a_1 \rightarrow a_5) \rightarrow a_4$ puisque la classe a_1a_5 associée à la règle $(a_1 \rightarrow a_5)$ n'est pas présente dans H_A . L'ensemble des R-règles associé à H_A est

$$H_{A \rightarrow} = \{a_1, a_2, a_3, a_4, a_5, a_2 \rightarrow a_3, a_5 \rightarrow a_4, (a_1 \rightarrow (a_5 \rightarrow a_4))\}$$

où les simples attributs sont ajoutés par complétude.

2.1 R-règles

Intuitivement, les R-règles sont une extension des règles binaires aux règles de règles.

Définition 2.1. Les R-règles de degré 0 sont composées des attributs de A . Les R-règles de degré 1 sont des règles binaires $a_i \rightarrow a_j$ entre deux attributs. Une R-règle, notée $R' \rightarrow R''$, entre deux R-règles R' et R'' de degrés respectifs j et k est une règle de degré $i = j + k + 1$. L'ensemble de toutes les R-règles possibles sur A est noté $R(A)$.

Par exemple, $a_2 \rightarrow a_3$ est une R-règle de degré 1 et $(a_1 \rightarrow (a_5 \rightarrow a_4))$ une R-règle de degré 2.

Les R-règles permettent d'exprimer différents degrés d'abstraction dans la description des relations implicatives. Une conjonction de R-règles de degré 0 est une description d'individus. Une R-règle de degré 1 est une implication entre descripteurs. Une R-règle de degré supérieur représente une implication entre implications. Par conséquent, leur interprétation se décline selon trois cas :

- 1- lorsque $R \rightarrow a_i$ alors a_i peut être interprété comme une conséquence de R – sachant évidemment qu'il s'agit d'une quasi-implication validée dans un contexte statistique et non d'une implication logique. De plus, bien que nous soyons dans un cadre statistique, l'intuition peut être ici guidée par les algèbres de Heyting pour lesquelles une implication $(a \Rightarrow b) \Rightarrow c$ est équivalente à $(a \text{ AND } b) \Rightarrow c$.
- 2- la R-règle $a_i \rightarrow R$ signifie que R peut se déduire de l'observation de a_i ;
- 3- la R-règle $R' \rightarrow R''$ signifie que la propriété R'' est un corollaire de la propriété R' précédemment démontrée .

2.2 Hiérarchies orientées

Notons $P(A)$ l'ensemble de toutes les k -permutations sur A pour $k=1$ à p . Les éléments de $P(A)$ peuvent être considérés comme des mots sur l'alphabet A composés de caractères différents. Les classes d'une hiérarchie orientée sont donc des mots particuliers. Nous considérons sur ces mots l'ordre de lecture de gauche à droite $<_L$. Par exemple, l'ordre sur $a_1a_3a_2$ est défini par $a_1 <_L a_3 <_L a_2$.



Comme nous l'avons indiqué dans notre exemple introductif, nous avons besoin pour construire une hiérarchie orientée de définir des opérateurs permettant de combiner et de comparer les classes entre elles. Reprenant pour simplifier le vocabulaire de la théorie des ensembles nous définissons les trois opérateurs suivants :

- *Intersection.* L'intersection $C' \cap C''$ de deux mots C' et C'' de $P(A)$ est le plus long mot commun à C' et C'' . En cas d'égalité, nous retenons le mot de C' commençant le plus à gauche selon $<_L$. Par exemple, si $C' = a_1a_3a_4 a_2$ et $C'' = a_7a_3a_4$ alors $C' \cap C'' = a_3a_4$ et si $C' = a_1a_2a_3 a_4$ et $C'' = a_3a_4 a_1a_2$ alors $C' \cap C'' = a_1a_2$.
- *Union.* L'union $C' \cup C''$ de deux mots C' et C'' distincts lettre à lettre ($C' \cap C'' = \emptyset$) est la concaténation de C' et C'' avec C' devant selon $<_L$. Si $C' = a_1a_3a_4a_2$ et $C'' = a_7a_8a_9$ alors $C' \cup C'' = a_1a_3a_4 a_2 a_7a_8a_9$.
- *Différence.* Pour tout triplet C, C', C'' de $P(A)$ tel que $C = C' \cup C''$, alors la différence $C - C'$ entre C et C' est égale à C'' , et la différence $C - C''$ entre C et C'' est égale à C' . Si $C = a_1a_2a_3$, $C' = a_1a_2$ et $C'' = a_3$ alors $C - C' = a_3$ et $C - C'' = a_1a_2$.

Définition 2.2. Une hiérarchie orientée H_A est un sous-ensemble de permutations de $P(A)$ qui vérifient les trois conditions ci-dessous :

1. H_A contient les attributs de A , appelés classes élémentaires ;
2. pour chaque classe C' et C'' de H_A , on a $C' \cap C'' \in \{ \emptyset, C', C'' \}$;
3. chaque classe non élémentaire C de H_A admet une décomposition unique en classes de H_A : il existe une unique paire C', C'' de classes de H_A telle que $C = C' \cup C''$.

Prise isolément l'interprétation d'une classe d'une hiérarchie orientée en R-règle est ambiguë ; par exemple, si on analyse la classe $a_1a_5a_4$ seule on ne sait pas si elle est associée à la R-règle $a_1 \rightarrow (a_5 \rightarrow a_4)$ ou $(a_1 \rightarrow a_5) \rightarrow a_4$. Cependant, la prise en compte de la hiérarchie orientée dans son intégralité permet de lever cette ambiguïté : $a_1a_5a_4$ est associée à la R-règle $a_1 \rightarrow (a_5 \rightarrow a_4)$ si les classes a_1 et a_5a_4 correspondant respectivement à la règle élémentaire a_1 et la règle $a_5 \rightarrow a_4$ sont dans H_A .

Proposition 2.1. (Gras et Kuntz, 2005). Chaque classe non élémentaire C d'une hiérarchie orientée H_A peut être associée à une unique R-règle de $R(A)$ notée C_{\rightarrow} .

La preuve repose sur une décomposition récursive de C en sous-classes appartenant à H_A . L'ensemble des classes de H_A peut alors être représenté par un arbre binaire orienté indicé :

- chaque classe élémentaire est une feuille de l'arbre ;
- chaque R-règle non élémentaire est représentée par un nœud de l'arbre qui est représenté ici par une flèche indiquant le sens de la relation implicative ; la hauteur $h(C) \in R^+$ d'un nœud associé à C satisfait la condition suivante : pour chaque nœud C' de H_A tel que $C' \cap C'' = C'$ (relation d'inclusion entre mots) alors $h(C) < h(C')$.

3 Validation des R-règles d'une hiérarchie orientée

Sur un corpus réel, l'objectif est de découvrir les R-règles qui sont statistiquement significatives selon un critère de qualité sur $P(A)$. Dans le cadre de l'analyse statistique implicative, le critère choisi repose sur l'intensité d'implication. Nous rappelons brièvement cette mesure définie pour les règles binaires simples (R-règles de degré 1) et l'étendons aux R-règles.



3.1 Intensité d'implication

Rappelons brièvement que, dans le cadre de l'analyse statistique implicative, il s'agit pour évaluer la qualité d'une règle $a_i \rightarrow a_j$ de modéliser la surprise suscitée par cette règle par rapport au comportement attendu (sous l'hypothèse d'indépendance deux à deux des variables). En d'autres termes, si $n(a_i \wedge \text{non } a_j)$ est le nombre de contre-exemples de la règle et $X(a_i \wedge \text{non } a_j)$ la variable aléatoire associée dans un modèle aléatoire, la mesure de la qualité de la règle est une fonction de la probabilité de l'écart entre $n(a_i \wedge \text{non } a_j)$ et $X(a_i \wedge \text{non } a_j)$. Notons $n(a_i)$ (resp. $n(a_j)$) le nombre d'occurrences de a_i (resp. a_j). Et, supposons que l'on tire aléatoirement dans I deux sous-ensembles avec $n(a_i)$ et $n(a_j)$ éléments ; on considère alors comme variable aléatoire $X(a_i \wedge \text{non } a_j)$ le nombre de contre-exemples dans ce tirage.

Définition 3.1. L'intensité d'implication de la règle $a_i \rightarrow a_j$ est définie par

$$\varphi(a_i, a_j) = 1 - \Pr(X(a_i \wedge \text{non } a_j) \leq n(a_i \wedge \text{non } a_j)) \text{ si } n(a_i) \neq n_j, \quad (1)$$

$$\text{et } \varphi(a_i, a_j) = 0 \text{ sinon.}$$

Pour un seuil de signification α donné, la règle $a_i \rightarrow a_j$ est retenue si $\varphi(a_i, a_j) \geq 1 - \alpha$.

La distribution de probabilité de $X(a_i \wedge \text{non } a_j)$ dépend du mode de tirage choisi. En pratique, on considère un tirage avec remise et nous utilisons pour calculer la loi de $X(a_i \wedge \text{non } a_j)$ une approximation par une loi normale.

3.2 Cohésion

La construction d'une hiérarchie orientée H_A dépend étroitement du critère d'agrégation choisi sur $P(A)$. Il s'agit de découvrir des R-règles $R' \rightarrow R''$ avec des relations d'implication fortes entre les attributs de R' et ceux de R'' .

Par exemple, il semble naturel de construire la R-règle $(a_1 \rightarrow a_2) \rightarrow (a_3 \rightarrow a_4)$ si les relations d'implications $a_1 \rightarrow a_3$, $a_1 \rightarrow a_4$, $a_2 \rightarrow a_3$ et $a_2 \rightarrow a_4$ sont suffisamment fortes. Ainsi, l'indice que nous avons défini pour quantifier ce que nous appelons la « cohésion » d'une R-règle est défini par une moyenne géométrique des différentes cohésions des règles mises en jeu.

Définition 3.2. Soit une R-règle $R' \rightarrow R''$, où R' et R'' sont respectivement associées aux permutations a'_1, a'_2, \dots, a'_k et $a''_1, a''_2, \dots, a''_h$. La cohésion $c(R)$ de R est définie par

$$c(R) = \left(c(R') \cdot c(R'') \cdot \prod_{i=1, k; k=1, h} c(a'_i, a''_j) \right)^{2/r(r-1)} \quad (2)$$

où $c(R') = \left(\prod_{i=1, k-1; j=2, k} c(a'_i, a'_j) \right)$ et $c(R'') = \left(\prod_{i=1, h-1; j=2, h} c(a''_i, a''_j) \right)$ avec $r = k + h$.

La cohésion $c(a_i, a_j)$ d'une R-règle de degré 1 est mesurée par un contraste entre la valeur de l'implication observée et le désordre associé à une expérience aléatoire que nous mesurons par une entropie ; la cohésion $c(a_i, a_j)$ est définie par

$$c(a_i, a_j) = (1 - (-p \log p - (1-p) \log(1-p))^{1/2}) \text{ si } p = \varphi(a_i, a_j) > 0.5 ; 0 \text{ sinon} \quad (3)$$



La valeur seuil 0.5 est atteinte par φ lorsque le nombre de contre-exemples observés est égal au nombre de contre-exemples attendus dans l'expérience aléatoire ; ainsi, lorsque φ est inférieure à 0.5 la surprise de l'implication est perdue d'où l'annulation de la cohésion. La cohésion d'une R-règle de degré > 1 peut se calculer en remplaçant $c(a'_i, a''_j)$ par la formule (3) dans la définition (2).

4 Construction d'une hiérarchie orientée

La construction d'une hiérarchie orientée est itérative et est obtenue par une méthode ascendante similaire à la démarche classique de la classification hiérarchique. Elle est initialisée au niveau 0 par les attributs (R-règles de degré 0). Puis, à chaque niveau h_i , une nouvelle règle est construite en déduction d'une union –au sens défini au paragraphe 2- de règles construites aux niveaux précédents. L'union retenue est celle qui maximise la valeur de l'indice de cohésion.

Ainsi,

- au niveau h_1 , la R-règle règle $a_i \rightarrow a_j$ construite est l'union des deux attributs de A qui maximisent la cohésion $c(a_i, a_j)$;
- au niveau h_2 , la R-règle construite est composée soit de deux attributs qui n'ont pas encore été agrégés, soit de la R-règle construite au niveau h_1 et d'un attribut non encore agrégé. La R-règle sélectionnée est celle qui maximise la cohésion ;
- au niveau h_3 , la R-règle construite peut être de trois types : une R-règle de degré 1 composée de deux attributs, une R-règle de degré 2 composée d'une R-règle construite en h_1 ou h_2 et d'un attribut, ou une R-règle de degré 3 composée de deux R-règles de degré 1 construites en h_1 et h_2 . Parmi les choix, la R-règle sélectionnée est celle qui maximise la cohésion ;
- Le processus est itéré jusqu'à ce que les cohésions de chacune des R-règles potentielles soient nulles.

Nous renvoyons à (Gras et Kuntz 2005) pour une description formelle de l'algorithme et de sa complexité. Dans l'exemple de la figure 1, l'agrégation de a_1 et $a_5 \rightarrow a_4$ au niveau h_3 signifie que la R-règle $a_1 \rightarrow (a_5 \rightarrow a_4)$ a la plus grande cohésion parmi les R-règles de l'ensemble des R-règles potentielles suivant

$$\{ a_1 \rightarrow (a_2 \rightarrow a_3), a_2 \rightarrow (a_3 \rightarrow a_1), (a_5 \rightarrow a_4) \rightarrow (a_2 \rightarrow a_3), \dots \}$$

L'algorithme s'arrête au niveau h_3 si les R-règles $(a_1 \rightarrow (a_4 \rightarrow a_5)) \rightarrow (a_2 \rightarrow a_3)$ et $(a_2 \rightarrow a_3) \rightarrow (a_1 \rightarrow (a_4 \rightarrow a_5))$ ont une cohésion nulle.

Définition 4.1. Pour chaque classe C de H_A , on note $c(C)$ la cohésion de la R-règle C_{\rightarrow} associée à C . La hauteur h de C est définie par $h(C) = 1 - c(C)$ si C n'est pas une classe élémentaire. On pose $h(C) = 0$ sinon.

On peut construire à partir de la hauteur h une dissimilarité u sur $A \times A$ de la façon suivante :

$$\begin{aligned} u(a_i, a_j) &= 1 \text{ si } a_i \text{ et } a_j \text{ ne sont pas agrégés dans } H_A ; \\ u(a_i, a_j) &= h(C_{ij}) \text{ sinon,} \\ &\text{où } C_{ij} \text{ est la plus petite classe de } H_A \text{ qui contient à la fois } a_i \text{ et } a_j. \end{aligned}$$

Comme en classification hiérarchique classique, on peut vérifier que la dissimilarité u ainsi construite est positive, symétrique et vérifie l'inégalité ultramétrique :

$$u(a_i, a_j) \leq \text{Max} (u(a_i, a_j), u(a_i, a_k)) \text{ pour tout } a_i, a_j, a_k \text{ de } A$$



Cela nous permet de justifier *a posteriori*, d'après le théorème d'équivalence entre une hiérarchie indicée et une ultramétrie de Johnson-Benzécri (e.g. Benzécri 1973), le choix du terme hiérarchie.

5 Significativité des niveaux

Comme en classification hiérarchique classique, étant donné la multiplicité des niveaux de la hiérarchie orientée, il est nécessaire de dégager ceux qui sont les plus pertinents par rapport à l'intention classificatrice de l'utilisateur et eu égard aux critères de construction choisis. Cette problématique peut être envisagée selon deux points de vue complémentaires : un point de vue global qui cherche à quantifier la qualité de chacune des partitions associées à chaque niveau de la hiérarchie, et un point de vue local qui se focalise sur la qualité des R-règles –assimilables dans une première approche à des classes- construites à chaque niveau.

Le premier point de vue, inspiré très étroitement d'une démarche proposée par I. Lerman, 1981, a été traité par R. Gras (Gras et Ratsimba-Rajohn 1996b). Le critère de significativité d'un niveau de la hiérarchie orientée est défini à partir d'une préordonnance ω induite par l'indice de cohésion. Il s'agit alors de comparer l'ensemble des couples de couples de $A \times A$ qui respectent la préordonnance initiale ω avec celui des couples de couples qui respecteraient une préordonnance aléatoire ω^* dans l'ensemble de toutes les préordonnances de même cardinal que ω muni d'une probabilité uniforme.

Le second point de vue ne porte plus sur le préordre sur l'ensemble des couples d'attributs mais sur celui défini sur les couples d'attributs « agrégés » à un même niveau de la hiérarchie orientée pour former une R-règle (Gras et al 2004). Il s'agit de comparer le nombre d'inversions entre l'ordre observé dans la classe et celui induit du modèle statistique associé à l'intensité d'implication, au nombre d'inversions attendu avec un ordre aléatoire sur un ensemble de même cardinal.

5.1 Critère basé sur une préordonnance

La cohésion définit sur l'ensemble des couples de $A \times A$ une relation d'ordre total $\omega : (a_i, a_j) \omega (a_k, a_l)$ si et seulement si $c(a_i, a_j) < c(a_k, a_l)$. Et, à un niveau quelconque h de la hiérarchie il existe une partition de $A \times A$ en deux ensembles : l'ensemble R_h des couples qui ont été agrégés au niveau h ou à des niveaux précédents, et l'ensemble $S_h = A \times A - R_h$ des couples d'attributs « séparés ». Par conséquent, l'ensemble des couples qui vérifient le préordre au niveau h est défini par $G(\omega) \cap (S_h \times R_h)$ où $G(\omega) = \{(C', C'') \in A \times A, c(C') < c(C'')\}$. L'adéquation entre $G(\omega)$ et $(S_h \times R_h)$ est déduite d'une comparaison entre le cardinal observé de l'ensemble de ces couples et celui déduit d'une préordonnance aléatoire. Considérons la variable aléatoire $Z = G(\omega^*) \cap (S_h \times R_h)$ où ω^* est une préordonnance aléatoire dans l'ensemble des préordonnances de même cardinal que ω muni d'une probabilité uniforme. L'adéquation entre $G(\omega)$ et $(S_h \times R_h)$ est mesuré par l'indice de similarité $s(\omega, h)$ suivant (Lerman 1981) :

$$s(\omega, h) = (\text{card}(G(\omega) \cap (S_h \times R_h)) - \mu) / \sigma \quad (4)$$

où μ et σ sont respectivement la moyenne et l'écart-type de Z .

Les niveaux privilégiés de la hiérarchie orientée sont ceux correspondant à un maximum local (meilleur que le niveau précédent et le niveau suivant) de $s(\omega, h)$.

5.2 Critère local



Une classe C de la hiérarchie orientée H_A formée au niveau k est considérée comme « cohérente » pour un seuil α , si il y a conformité ou quasi-conformité au seuil α entre l'ordre –ou le préordre- ω_0 dans lequel s'organise les attributs de C selon la cohésion et l'ordre –ou le préordre- théorique ω_i défini par leurs intensités d'implication mutuelles.

Pour évaluer précisément cette conformité, nous nous basons sur une propriété de l'intensité d'implication (Gras et Larher 1992) : si le nombre d'occurrences de a_i est inférieur au nombre d'occurrences de a_j , alors la qualité de $a_i \rightarrow a_j$ au sens de φ est meilleure que celle de sa réciproque $a_j \rightarrow a_i$. Ainsi, l'ordre théorique ω_i défini par les intensités d'implications mutuelles coïncide avec celui défini par les occurrences des attributs. Nous comparons la conformité entre ω_0 et ω_i avec celle entre un ordre aléatoire ω^* et ω_i . Nous mesurons la conformité par le nombre d'inversions entre les différents ordres : inv est le nombre d'inversions observées entre ω_0 et ω_i et Inv est le nombre d'inversions entre ω^* et ω_i . Le nombre d'inversions entre deux ordres est simplement défini ici par le nombre de paires d'attributs (a_i, a_j) telles que a_i est avant a_j dans le premier ordre et après dans le second.

Intuitivement cela signifie que, si α est petit, la conformité entre ω_0 et ω_i est invraisemblablement très grande puisqu'il paraît exceptionnel que le hasard « fasse mieux » que ce qui est observé.

Définition 5.1. La cohérence $o(C)$ d'une classe C d'une hiérarchie orientée est définie par la probabilité $Pr(Inv > inv)$.

Ainsi, plus le nombre d'inversions est faible, eu égard à la cardinalité de la classe, plus grande est la cohérence de la classe. De plus, pour un même nombre d'inversions observées pour deux classes C' et C'' , si la classe C' contient plus d'attributs que la classe C'' , la cohérence de C' est meilleure que celle de C'' .

Le calcul de la cohérence repose sur la détermination de la loi de Inv qui dépend du nombre d'attributs p . Notons que nous retrouvons cette variable aléatoire dans le calcul du coefficient de corrélation des rangs de Kendall. Nous proposons dans (Gras et al 2004) une formule de récurrence permettant de calculer facilement ses valeurs dans l'indice de cohérence.

Dans une perspective d'aide à l'interprétation nous cherchons à établir un critère de significativité d'une classe de la hiérarchie orientée et un indicateur permettant de sélectionner les niveaux les plus pertinents.

Afin de restituer l'information maximale relative à l'ensemble des classes constituées, la significativité doit intégrer deux paramètres majeurs : les cohésions des classes dont, par construction de H_A , les valeurs décroissent avec la croissance des niveaux de la hiérarchie, et les cohérences des classes qui peuvent croître ou décroître selon les niveaux en fonction de la probabilité associée à la variable aléatoire Inv eu égard aux inversions observées et à la taille de la classe. Nous proposons de façon *ad hoc* une mesure qui satisfait les quatre contraintes suivantes liées à la « sémantique » de la significativité :

- être une fonction de la cohérence et de la cohésion qui majore les valeurs de la cohérence ;
- conserver l'aspect probabiliste que possède la cohérence ;
- pondérer la cohérence, indice de « bon ordre » des attributs dans la classe selon l'implication par un facteur qui pourrait être qualifié d'affaiblissement de la cohésion qui vise selon les cas : (i) à prendre en compte favorablement le fait que la classe formée au niveau $k + 1$ ait une cohésion peu différente de la classe formée à niveau k , (ii) à prendre en compte défavorablement le fait que la différence étant élevée, cela affecte la crédibilité de la classe formée en $k + 1$ même si elle a une bonne cohésion ;
- diminuer la significativité d'une classe au niveau $k + 1$ qui bien qu'ayant une bonne cohérence a une cohésion qui décroît entre k et $k + 1$.

Définitions 5.2. L'indice co de cohésion-cohérence qui mesure la significativité de la classe C_{k+1} formée au niveau $k + 1$ est défini par



$$co(C_{k+1}) = \frac{c(C_{k+1})}{c(C_k)} \cdot o(C_{k+1}) \quad (5)$$

Et, par convention, $co(C_0) = 1$.

Un niveau k de la hiérarchie orientée H_A est significatif si il correspond à un maximum local de l'indice de cohésion-cohérence de la classe formée à ce niveau.

En effet, l'indice co n'étant pas une fonction monotone, il apparaît des maxima locaux correspondant d'une part à une meilleure adéquation, entre les restrictions à la classe formée à ce niveau, des préordres théoriques ω_i et contingents ω_b , d'autre part à une bonne cohésion.

Définitions 5.3. La qualité de l'ensemble des niveaux h , $0 \leq h \leq k$, est définie par

$$q_k(H_A) = \left(\prod_{h=1}^k co(C_h) \right) \quad (6)$$

où C_h désigne la classe formée au niveau h .

La hiérarchie orientée H_A est significative au niveau h si sa qualité $q_h(H_A)$ admet un minimum local.

6 Conclusion

L'analyse des R-règles a démontré sa pertinence dans différents champs applicatifs (e.g. didactique (Bodin et Gras 1999), ressources humaines (Peter et al 2001)). L'expérience nous a montré que les R-règles associées à des classes de faible degré (degré < 5) sont généralement facilement interprétables et peuvent apporter un supplément d'information significatif. L'analyse permet notamment de compléter la première organisation fournie par le graphe implicatif.

En revanche, dès que les règles se complexifient, il peut devenir délicat de leur associer une sémantique claire et facilement exploitable. Si l'analyse des R-règles en tant que telles trouve alors certaines limites, en revanche, la construction de la hiérarchie implicative permet d'associer à l'ensemble des attributs une proposition de structuration générale. Ainsi, les hiérarchies orientées offrent, pour différents niveaux de granularité, un partitionnement préalable de l'ensemble des attributs en classes dont les éléments entretiennent entre eux des relations hiérarchiques.

En complément d'un critère de sélection des niveaux les plus significatifs, se pose la question des contributions respectives des individus ou des classes d'individus à la constitution des classes. Selon un schéma inspiré de l'analyse des correspondances, une première piste explorée consiste à plonger l'ensemble des individus dans un espace métrique adapté (Gras et al 2001). Des réflexions sont en cours pour la validation à la fois théorique et expérimentale de cette approche.

Références

- Benzécri J.-P. (1973), L'analyse des données (vol. 1) : Taxonomie, Dunod, Paris.
- Bodin A. et Gras R. (1999). Analyse du préquestionnaire enseignants avant EVAPM-terminales, Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement Public, n° 425, pp. 772-786.
- Domenach F. et Leclerc B. (2003), Closure systems, implicational systems, overhanging relations and the case of hierarchical classification, à paraître.
- Gras R. (1979), Contribution à l'analyse expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques. Thèse d'Etat, Université de Rennes 1.
- Gras R. et Lahrer A (1992), L'implication statistique, une nouvelle méthode d'analyse de données. Mathématique, Informatique et Sciences Humaines, 120, pp 5-31.



- Gras R. et al (1996a), L'implication statistique – Nouvelle méthode exploratoire de données, La Pensée Sauvage éditions, France.
- Gras R. et Ratsimba-Rajohn H. (1996b), Analyse non symétrique de données par l'implication statistique. RAIRO-Recherche Opérationnelle, 30(3), pp. 217-232.
- Gras R (1997), Nœuds et niveaux significatifs en analyse statistique implicative. Prépublication 97-32, Institut de Recherche de Mathématiques de Rennes.
- Gras R., Kuntz P. et Briand H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, Mathématiques et Sciences Humaines, n° 154-155, pp. 9-29.
- Gras R. et Kuntz P. (2003), Hiérarchie orientée de règles généralisées en analyse implicative. Extraction des Connaissances et Apprentissage, 17(1), pp. 145-157.
- Gras R. et Kuntz P. (2005), Discovering R-rules with a directed hierarchy., Soft Computing Journal, à paraître.
- Gras R., Kuntz P. et Régnier J.-C. (2004), Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative, Numéro spécial Classification, Revue des Nouvelles Technologies de l'Information, Cépaduès.
- Horschka P. et Klögsen W. (1991), A support system for interpreting statistical data. Knowledge Discovery in Databases, AAAI Press, pp. 325-345.
- Lehn R. (2000), Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans une base de données. Thèse de doctorat, Université de Nantes.
- Lent B., Swami A.N. , et Widow J. (1997), Clustering association rules. Proc. of the 13th Int. Conf. on Data Engineering, pp. 220-231.
- Lerman I.C. (1981), Classification et analyse ordinale des données. Dunod, Paris.
- Peter P. R. Gras, Philippé J. et Baquédano S. (2001), L'analyse implicative pour l'étude d'un questionnaire de personnalité, Extraction et Gestion des Connaissances, vol. 1, n° 1, pp. 251-258.
- Rostam H. (1981), Construction automatique et évaluation d'un graphe d'implication issu de données binaires dans le cadre de la didactique des mathématiques, Thèse de doctorat, Université de Rennes I.
- Toivonen H. , Klementtinen M., Ronkainen P., Hätonen K. et Manila H. (1995), Pruning and grouping of discovered association rules. Workshop notes of the ECML Workshop on Statistics, Machine Learning and Knowledge Discovering in Databases, pp. 47-52.

Summary

This communication is a synthesis of recent results in ISA which extend the classical notion of quasi-implication (« when a_i is present then usually a_j is also present») to R-rules (rules of rules), the premisses and the conclusions of which can be rules themselves. We define a new measure for validating the statistical significance of these R-rules, and develop the concept of “directed hierarchy” to structure a set of R-rules. In order to make the interpretation of this new structuration easier, we propose a criterium for selecting the most relevant levels of the hierarchy like in classical hierarchical classification.