Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative

Régis Gras*, Pascale Kuntz*
Jean-Claude Régnier**

*Laboratoire d'Informatique de Nantes-Atlantique FRE 2729
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie – BP 60601
44306 Nantes cedex 3
regisgra@club-internet.fr
pascale.kuntz@polytech.univ-nantes.fr
**EA 3727 Savoirs, Diversité et Professionnalisation
86, rue Pasteur
69365 Lyon cedex 07
Jean-claude.regnier@univ-lyon2.fr

Résumé. Dans le cadre de l'analyse statistique implicative développée à l'origine par R. Gras, nous avons proposé le modèle de « hiérarchie orientée » pour structurer des règles partielles de type $a \rightarrow b$ et des règles de règles, appelées *R*-règles, issues d'un corpus de données décrites par des attributs binaires. Dans cet article, nous proposons un nouveau critère de significativité[1] des niveaux de la hiérarchie orientée basé sur des comparaisons de préordres. Une application à un questionnaire d'opinions illustre l'intérêt de la démarche.

1. Introduction

Introduites en Extraction des Connaissances dans les Données au début des années *90* par R. Agrawal, [Agrawal et al., 1993], pour exprimer simplement des tendances implicatives $A_i \rightarrow A_j$ entre des sous-ensembles d'attributs A_i et A_j d'une table relationnelle, les règles d'association ont rapidement connu une utilisation intensive. Contrairement aux approches initiales de l'analyse combinatoire des données et de la classification conceptuelle, la condition d'inclusion $I(A_i) \subset I(A_j)$ entre le sous-ensemble d'individus $I(A_i)$ décrit par A_i et le sous-ensemble $I(A_j)$ décrit par A_j est ici relaxée, et l'on considère que la tendance générale à posséder A_j quand on a A_i n'est pas rejetée par la situation, fréquente pour des données réelles, où l'on observe quelques rares contre-exemples.

De nombreux algorithmes, dont le plus célèbre est certainement *Apriori*, ont été proposés dans la littérature. Cependant, il est bien connu dans la pratique qu'ils engendrent un nombre prohibitif de règles pour une analyse directe *in extenso*. Il est devenu alors nécessaire de

[1] Dans ce texte, le mot « significativité » aura un sens plus général que statistique. Il exprimera : « …qui est révélateur d'un phénomène d'intérêt sémantique majeur », même si la référence à une échelle de probabilité restera généralement présente.

présenter les règles sous une forme organisée, la structuration pouvant être partie intégrante d'un processus de fouille, ou intervenir en seconde étape sur l'ensemble *S* des règles pré-sélectionnées par un algorithme automatique.

1.1 Structuration des quasi-implications

Une première approche, couramment employée pour ses facilités de mise en œuvre et d'interprétation, consiste à représenter les relations sur *S* par des graphes orientés. Dans le modèle le plus simple, les sommets du graphe sont les prémisses et les conclusions, les arcs représentent les quasi-implications. Si un tel graphe permet, sans légende, d'appréhender les classes de la relation d'équivalence de base « avoir une prémisse ou une conclusion commune », il permet en revanche difficilement, hormis la transitivité quand elle existe, de déduire d'autres relations. D'autres modèles ont ainsi été proposés pour permettre des inférences déductives [Horschka et Klögsen, 1991 ; Lehn, 2000]. Une deuxième voie consiste à rechercher une structuration d'un autre ordre basée sur des modèles de classification. Sans connaissance préalable d'un modèle spécifiquement approprié, tel que par exemple le modèle hiérarchique associé à un système implicatif complet, la démarche classique consiste à classifier *S* en recherchant un sous-ensemble de l'ensemble de ses parties [Lent et al., 1997 ; Toivonen et al., 1995]. Les travaux publiés en ECD recherchent essentiellement des partitions déduites, par des approches classiques, de dissimilarités construites sur $S \times S$ ou sur le produit cartésien de l'ensemble des prémisses et des conclusions. Par la démarche souvent *ad hoc*, la structure ainsi obtenue est généralement difficile à interpréter. Prolongeant un travail initié par R. Gras, [Gras, 1979] puis dans [Gras et Larher, 1992], nous avons récemment développé une démarche alternative [Gras et Kuntz, 2003], 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 prémisse *R* et la conclusion *R'* peuvent être elles-mêmes des règles. On parle alors de *R*-règles. Le modèle proposé, appelé « hiérarchie orientée », est une extension du modèle hiérarchique classique, sur l'ensemble des parties de l'ensemble *A* des attributs, à un ensemble de règles : les niveaux de la hiérarchie orientée sont des règles ou des *R*-règles. Contrairement au modèle 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.

1.2 Significativité des niveaux de la hiérarchie orientée

Comme en classification hiérarchique classique, étant donnée 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 en fonction des 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.C. Lerman, [Lerman, 1981], a été traité dans [Gras et Ratsimba-Rajohn, 1996]. Le critère de significativité d'un niveau de la hiérarchie orientée est défini à partir d'une préordonnance Ω

induite par un indice sur $A \times A$, appelé indice de cohésion, défini pour valider la qualité implicative des *R*-règles. 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.

Dans cet article, nous nous focalisons sur le second point de vue. Ainsi, au lieu de nous intéresser au préordre sur l'ensemble des couples d'attributs, nous nous intéressons au préordre défini sur les couples d'attributs « agrégés » à un même niveau de la hiérarchie orientée pour former une *R*-règle. Nous comparons le nombre d'inversions entre l'ordre observé dans la classe et celui induit d'un modèle statistique, l'intensité d'implication, au nombre d'inversions attendu avec un ordre aléatoire sur un ensemble de même cardinal.

Dans le deuxième paragraphe de cet article, nous rappelons brièvement le cadre méthodologique dans lequel nous nous situons pour valider la qualité des règles et des *R*-règles, et pour construire les hiérarchies orientées. Dans le troisième paragraphe, nous définissons un premier critère local contribuant à établir la significativité, appelé « cohérence ». Sa mise en œuvre nécessitant la caractérisation de la loi suivie par une variable aléatoire donnant le nombre d'inversions dans un modèle aléatoire, nous consacrons le quatrième paragraphe à ce problème, proposons et démontrons une formule de récurrence et quelques propriétés asymptotiques. Le concept de cohésion-cohérence est défini, dans le cinquième paragraphe, pour permettre de déclarer la significativité d'un niveau de la hiérarchie. Une illustration, sur un jeu de données réelles issu d'une enquête auprès d'enseignants de mathématiques de Terminale sur leurs attentes relatives aux objectifs pédagogiques, est proposée dans la dernière partie.

2. Cadre méthodologique

Dans la suite, nous considérons un ensemble I d'individus décrits par un ensemble fini $A = \{a_1, a_2, \ldots.\}$ de m attributs binaires. On note Ω_A l'ensemble de toutes les k-permutations de A, pour $k = 1$ à m, et l'ordre de lecture sur les attributs d'une k-permutation est noté $<$.

Définition 2.1. Une *hiérarchie orientée* H_A sur A est un ensemble d'éléments de Ω_A, appelés *classes*, vérifiant les trois conditions suivantes :

1. H_A contient tous les attributs de A, appelés classes élémentaires ;
2. Pour chaque couple C', C'' de classes de H_A, on a $C' \tilde{\cap} C'' = \{\emptyset, C', C''\}$, où l' « intersection » $\tilde{\cap}$ entre deux séquences de Ω_A est définie comme étant la plus grande sous-séquence d'attributs contigus communs à C' et C'' (par exemple, $a_1a_3a_4a_2 \tilde{\cap} a_4a_2 = a_4a_2$) ; en cas d'égalité on retient la première, selon $<$, sous-séquence de C' ;
3. Pour toute classe non élémentaire C de H_A, il existe un unique couple C', C'' tel que $C = C' \tilde{\cup} C''$, où l' « union » $\tilde{\cup}$ de deux séquences disjointes de Ω_A est définie par la concaténation de C' et C'' selon l'ordre $<$ (par exemple, $a_1a_3a_4a_2 \tilde{\cup} a_7a_3a_4 = a_1a_3a_4a_2\, a_7a_3a_4$).

Par exemple, $H_A = \{a_1, a_2, a_3, a_4, a_5, a_2a_3, a_5a_4, a_1a_5a_4\}$ est une hiérarchie orientée sur $A = \{a_1, a_2, a_3, a_4, a_5\}$. Elle peut être traduite par un arbre dont les nœuds représentent des relations d'implications entre les attributs de A (voir une illustration fig.1 page 11) ; ces

relations peuvent être des quasi-implications simples telles que $a_2 \rightarrow a_3$, ou bien des *R*-règles telles que par exemple $a_1 \rightarrow (a_5 \rightarrow a_4)$.

Définition 2.2. Les *R-règles* de degré 0 sont les attributs de *A*, considérés implicitement de la forme $a_i \rightarrow a_i$. Les *R*-règles de degré 1 sont les quasi-implications simples de la forme $a_i \rightarrow a_j$. Une *R*-règle de degré *i*, $1 < i \leq p$, de la forme $R' \rightarrow R''$ entre deux R-règles *R'* et *R''* de degrés respectifs *j* et *k* vérifie $j + k = i - 1$.

Chaque classe *C* d'une hiérarchie orientée peut être associée à une unique *R*-règle [Gras, Kuntz et Briand, 2003a], ce qui facilite son interprétation. Ainsi, la hiérarchie H_A de l'exemple ci-dessus peut être associée à un ensemble unique de *R*-règles $\vec{H}_A = \{a_1, a_2, a_3, a_4, a_5, a_2 \rightarrow a_3, a_5 \rightarrow a_4, a_1 \rightarrow (a_5 \rightarrow a_4)\}$. Intuitivement, on voit bien ici qu'une *R*-règle de degré >0 construite à un niveau *k* résulte de l'« agrégation » de *R*-règles précédemment construites à des niveaux inférieurs ; par exemple, la *R*-règle $a_1 \rightarrow (a_5 \rightarrow a_4)$ de degré *2* associée à la classe $a_1a_5a_4$ résulte de l'agrégation de a_1 et de $a_5 \rightarrow a_4$ qui est associée à la classe a_5a_4.

La construction d'une hiérarchie orientée H_A dépend étroitement du critère d'agrégation choisi sur Ω_A. Il s'agit de découvrir des *R*-règles $R' \rightarrow R''$ avec des relations d'implication forte 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'implication $a_1 \rightarrow a_3$, $a_1 \rightarrow a_4$, $a_2 \rightarrow a_3$ et $a_2 \rightarrow a_4$ sont suffisamment fortes. Ainsi, l'indice *c* que nous avons défini pour quantifier la « cohésion » d'une *R*-règle $R' \rightarrow R''$, où *R'* et *R''* sont respectivement associées aux permutations $a'_1, a'_2, \ldots, a'_k$ et $a''_1, a''_2, \ldots, a''_k$, est de la forme suivante, où r=k+h :

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

Pour calculer *c*, nous nous sommes placés dans le cadre de l'analyse statistique implicative. Rappelons brièvement, que dans ce cadre, il s'agit, pour évaluer la qualité d'une quasi-implication $a_i \rightarrow a_j$, de modéliser la surprise suscitée par cette règle par rapport au comportement attendu sous l'hypothèse d'indépendance entre a_i et a_j ; en d'autres termes, si $n_{a_i \wedge \neg a_j}$ est le nombre de contre-exemples de la règle et $X_{a_i \wedge \neg a_j}$ la variable aléatoire associée dans un modèle aléatoire sous la même hypothèse, la mesure de la qualité de la règle est une fonction de la probabilité de l'écart entre $n_{a_i \wedge \neg a_j}$ et $X_{a_i \wedge \neg a_j}$. 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 respectivement n_{a_i} et n_{a_j} éléments ; on considère alors comme variable aléatoire $X_{a_i \wedge \neg a_j}$ le nombre de contre-exemples dans ce tirage.

Définition 2.3. L'*intensité d'implication* de la règle $a_i \rightarrow a_j$ est définie par
$\varphi(a_i, a_j) = 1 - \Pr(X_{a_i \wedge \neg a_j} \leq n_{a_i \wedge \neg a_j})$ si $n_{a_j} \neq n_l$, et $\varphi(a_i, a_j) = 0$ sinon (**2**)

En pratique, nous utilisons pour calculer la loi de $X_{a_i \wedge \neg a_j}$ une approximation par une loi normale.

Ainsi, 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

$$c(a_i, a_j) = (1 - (- p \log_2 p - (1 - p) \log_2 (1 - p))^2)^{1/2} \text{ si } p = \varphi(a_i, a_j) > 0.5 \text{ ; } 0 \text{ sinon} \qquad (3)$$

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 (2).

La construction de la hiérarchie orientée est itérative et obtenue par une méthode ascendante. Elle est initialisée, au niveau *0*, par les attributs. Puis, à chaque niveau on construit une nouvelle classe qui est une union au sens de la définition 2.1. de deux classes construites à des niveaux précédents qui maximisent la cohésion.

3. Critère de cohérence des niveaux

Une classe C de la hiérarchie orientée H_A formée au niveau k est considérée comme *cohérente* pour un seuil α, s'il y a conformité ou quasi-conformité au seuil α entre l'ordre – ou le préordre- ω_0 dans lequel s'organisent les attributs de C selon la cohésion et l'ordre –ou le préordre- théorique ω_t 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_t défini par les intensités d'implication mutuelles coïncide avec celui défini par les occurrences des attributs. Nous comparons la conformité entre ω_0 et ω_t avec celle entre un ordre aléatoire ω^* et ω_t. Nous mesurons la conformité par le nombre d'inversions entre les différents ordres : i est le nombre d'inversions observées entre ω_0 et ω_t et I est le nombre d'inversions entre ω^* et ω_t. 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_t est invraisemblablement très grande puisqu'il paraît exceptionnel que le hasard « fasse mieux » que ce qui est observé.

Définition 3.1. La *cohérence o(C)* d'une classe C d'une hiérarchie orientée est définie par la probabilité $Pr(I > i)$.

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''*.

Exemple 3.1. Considérons une classe *C* d'une hiérarchie orientée H_A constituée de cinq attributs, a_i, i =1 à 5 structurée selon l'ordre $\omega_0 = \{a_1, a_4, a_3, a_2, a_5\}$. On suppose d'autre part que leurs occurrences sont telles que $n_{a_1} < n_{a_2} < \ldots < n_{a_5}$, ce qui induit selon la propriété rappelée ci-dessus, un ordre théorique $\omega_t = \{a_1, a_2, a_3, a_4, a_5\}$ pour les intensités d'implication. On vérifie aisément que le nombre d'inversions *i* entre ω_0 et ω_t est *3* (échanges de a_3 et a_2, a_4 et a_3, a_4 et a_2). Afin d'évaluer la cohérence de la classe, il faut déterminer la loi de la variable *I*. Pour *5* attributs on peut obtenir pas à pas la distribution en énumérant les cas où chacun des attributs est minimal dans l'ordre ω^* (tableau 1). Ici, on a $Pr(I > i) = Pr(I > 3) = 91/120$. (soit environ ¾).

Nombre d'inversions	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
a_1 minimal	1	3	5	6	5	3	1	0	0	0	0
a_2 minimal	0	1	3	5	6	5	3	1	0	0	0
a_3 minimal	0	0	1	3	5	6	5	3	1	0	0
a_4 minimal	0	0	0	1	3	5	6	5	3	1	0
a_5 minimal	0	0	0	0	1	3	5	6	5	3	1
Total des permutations	*1*	*4*	*9*	*15*	*20*	*22*	*20*	*15*	*9*	*4*	*1*

TAB 1 – *Détermination de la distribution de I dans l'exemple*

D'une façon générale, la mise en œuvre de la cohérence définie en 3.1. nécessite de déterminer la loi de *I*, que nous noterons I_m dans la suite puisqu'elle dépend du nombre *m* d'attributs. Notons que le recours à la variable aléatoire I_m donnant le nombre d'inversions entre deux permutations est présent dans le calcul du coefficient de corrélation des rangs τ de Kendall qui peut effectivement s'écrire

$$1-\tau=\frac{4I_m}{m(m-1)} \qquad (4)$$

Cependant, à notre connaissance, la loi de I_m n'est ni explicitement donnée ni formalisée [Kendall et Stuart, 1991]. Nous proposons et établissons, dans la suite, une formule de récurrence permettant de calculer ses valeurs dans l'indice de cohérence.

4. Loi de la variable I_m

Sous l'hypothèse d'équiprobabilité des permutations, nous considérons la variable aléatoire $N(I_m(k))$ donnant le nombre total de permutations aléatoires correspondant à un nombre d'inversions avec ω_t égal à *k* pour un nombre d'attributs égal à *m*. Notons que l'on a trivialement $Pr(I_m = 0) = 1/m !$ puisque le nombre d'inversions est nul si et seulement si ω_t coïncide avec ω^*.

Proposition 4.1. Pour tout $k < m$, on a

$$N(I_m(k)) = \sum_{j=0}^{k} N\left(I_{m-1}(j)\right) \quad \textbf{(5)}$$

et, pour tout $k \geq m$, on a

$$N(I_m(k)) = \sum_{j=k-m+1}^{k} N\left(I_{m-1}(j)\right) \quad \textbf{(6)}$$

Preuve. Remarquons tout d'abord, que pour tout *k*, on a

$$N\left(I_m(k)\right) = \sum_{i=1}^{m} N\left(I_m(k); a_i\right) \quad \textbf{(7)}$$

où $N(I_m(k) ; a_i)$ est le nombre total de permutations lorsque l'attribut a_i est minimal dans l'ordre aléatoire ω^* et placé au *i*ème rang dans l'ordre théorique ω_t.

Supposons maintenant que $k < m$. Pour tout *i* de 1 à m, la place minimale de a_i dans entraîne $(i - 1)$ inversions ; les $k - (i - 1)$ autres inversions sont donc provoquées par les $m - 1$ autres attributs. Ainsi, pour tout *i* de 1 à $(k + 1)$ on a

$$N\left(I_m(k)\right) = \sum_{i=1}^{k+1} N\left(I_{m-1}(k-(i-1)\right) = \sum_{i=0}^{k} N\left(I_{m-1}(k-i)\right) = \sum_{j=0}^{k} N\left(I_{m-1}(j)\right) \quad \textbf{(8)}$$

et, pour tout *i* de $k + 2$ à *m*, $N(I_m(k) ; a_i) = 0$ puisque dans ce cas la variable a_i étant minimale il y a au moins $i - 1$ inversions auxquelles aucune permutation ne peut conduire.
La preuve de la formule de récurrence pour le cas $k \geq 1$ est basée sur un raisonnement similaire.

Les relations de la proposition 4.1. permettent de calculer les lois des variables I_m selon une formule de récurrence. En effet, connaissant la distribution de I_{m-1} on peut déterminer celle de I_m, et les valeurs initiales sont directement calculables. On vérifie aisément que $N(I_2(0) ; a_1) = 1$ (c'est la permutation de a_1 et a_2), $N(I_2(1) ; a_1) = 0$; $N(I_2(0) ; a_2) = 0$, $N(I_2(1) ; a_2) = 1$, d'où $N(I_2(0)) = 1$ et $N(I_2(1)) = 1$. On en déduit ainsi la loi de I_2 : $Pr(I_2 = 1) = Pr(I_2 = 0) = 0.5$, puis celle de I_3, etc ..

Proposition 4.2. Pour tout $k < m$, on a $N(I_m(k)) = N(I_{m-1}(k)) + N(I_m(k-1))$ et, pour tout $k \geq m$, on a $N(I_m(k)) = N(I_m(k-1)) + N(I_{m-1}(k)) - N(I_{m-1}(k-m))$.

Cette proposition se déduit d'arguments similaires à ceux employés dans la proposition précédente [Gras, 1997] et de la relation (7).
On peut ainsi déduire de façon récurrente les différentes valeurs de la loi de I_m utiles pour le calcul de la cohérence. En effet, pour $k = 1$ et $m > 1$, on déduit de la proposition 4.2. l'équation linéaire aux différences d'ordre 1 en *m* suivante : $N(I_m(1)) = N(I_m(0)) + N(I_{m-1}(1)) = 1 + N(I_{m-1}(1))$. D'où, $N(I_m(1)) - N(I_{m-1}(1)) = 1$ dont la solution avec second membre est $N(I_m(1)) = m - 1$. Par conséquent, rappelant que l'on a pour tout $m > 1$, $Pr(I_m = 0) = 1/m$! on obtient

$$Pr\left(I_m = 1\right) = \frac{m-1}{m!} \quad \textbf{(9)}$$

Ainsi, par exemple si $m = 2$, la cohérence associée à une situation sans inversion est $Pr(I_2 > 0) = Pr(I_2 = 1) = 0.5$.

De la même façon, pour $m > 2$, on obtient en utilisant d'une part, le résultat donnant $N(I_m(1))$ et d'autre part, le fait que $N(I_m(1 ; a_m)) = 0$, la relation $N(I_m(2)) = N(I_{m-1}(2)) + m - 1$. D'où, $N(I_m(2)) = m^2/2 - m/2 - 1$, et pour $m > 2$, on a donc

$$Pr(I_m = 2) = \frac{m^2 - m - 2}{2m!} \quad \textbf{(10)}$$

Puis, comme $N(I_m(3)) = N(I_{m-1}(3)) + m^2/2 - m/2 - 1$, on obtient pour $m > 3$,

$$Pr(I_m(3)) = \frac{m^2 - 7}{6(m-1)!} \quad \textbf{(11)}$$

Par exemple, on a $Pr(I_6 \leq 2) = 0.028$. Par suite, $Pr(I_6 > 2) = 0.972$ est la valeur de la cohérence d'une classe de *6* attributs dans laquelle on observerait *2* inversions entre l'ordre associé à la classe ω_0 et l'ordre théorique ω_t .

Remarque 4. 1. Pour une classe *C* réduite à un singleton, sa cohérence ne peut se déduire des relations précédentes. Nous posons dans ce cas $o(C) = 1/2$ du fait que l'absence d'inversion n'apporte aucune information puisqu'elle est nécessaire.

Remarque 4.2. Lorsque deux attributs d'une classe *C* ont le même nombre d'occurrences et ont donc ainsi le même rang dans le préordre ω_t , le nombre d'inversions qui leur sont relatives est calculé comme si les attributs avaient un rang distinct.

Proposition 4.3. L'espérance de la variable aléatoire I_m vaut $E(I_m) = m(m-1)/4$ et sa variance vaut $V(I_m) = m\,(m-1)\,(2m+5)\,/\,72$. Et, la loi de I_m converge vers une loi normale quand m tend vers l'infini.. L'espérance et la variance peuvent se déduire, par la relation (4), des résultats connus de la statistique τ dont l'espérance est $E(\tau) = 0$ et la variance $V(\tau) = 2\,(2m + 5)\,/\,9m\,(m-1)$. De plus, dans son article séminal [Kendall, 1938], Kendall expose les grandes lignes d'une démonstration permettant de déduire que, quand *m* tend vers l'infini, la variable

$$Z = \frac{\tau}{\sqrt{\frac{2(2m+5)}{9m(m-1)}}} \quad \textbf{(12)}$$

est asymptotiquement distribuée comme une variable de Laplace-Gauss centrée réduite, dont il donne la table, à partir de ses moments centrés d'ordre pair, sous l'hypothèse d'équiprobabilité des permutations.

$$\mu_{2k} \approx \frac{(2k)!}{2^k\,k!}(\mu_2)^k \quad \textbf{(13)}$$

Compte tenu de la relation entre I_m et τ, il est clair que la distribution de I_m est également asymptotiquement distribuée comme une variable gaussienne, dont l'approximation est acceptable à partir de m=10 [Siegel, 1956]. A titre d'exemple, voici une comparaison : $Pr(I_{20} = 50) \approx 0.0003197$ et $Pr(I_{20} > 50) \approx 0.9985028$; pour la loi *LG* on obtient $Pr[49.5 < LG(95 ; 15.41) < 50.5] \approx 0.0003647$ et $Pr(LG(95 ; 15.41) > 50.5) \approx 0.998059$

5. Vers un nouvel indice de significativité

A un niveau $k > 0$ donné, une hiérarchie orientée H_A présente plusieurs classes déjà formées –associées à des *R*-règles de degré > 0-, et éventuellement, quelques attributs non encore associés. Nous cherchons dans ce paragraphe à quantifier la significativité d'une classe, ainsi que la qualité de la hiérarchie à ce niveau.

Afin de restituer l'information maximale relative à l'ensemble des classes constituées, cette 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 I_m eu égard aux inversions observées et à la taille de la classe. Le concept que nous proposons (définition 5.1.) pour associer ces 2 paramètres satisfait les 4 contraintes suivantes liées à la « sémantique » de la significativité :

1. être fonction de la cohérence et de la cohésion majorant les valeurs de la cohérence ;
2. conserver l'aspect probabiliste que possède la cohérence ;
3. 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 et visant 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.
4. 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 5.1. L'indice *co* de *cohésion-cohérence* qui mesure la significativité de la classe C_{k+1} formée au niveau $k + 1$ est défini par

$$co(C_{k+1}) = \frac{c(C_{k+1})}{c(C_k)}.o(C_{k+1}) \quad (14)$$

Par convention, $co(C_0) = 1$. Un niveau k de la hiérarchie H_A est *significatif* s'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 théorique ω_t et contingent ω_0, et d'autre part à une bonne cohésion.

Définitions 5.2. La *qualité* de l'ensemble des niveaux h, $0 \le h \le k$, est définie par

$$q_k(H_A) = \left(\prod_{i=1}^{k} co(C_i) \right) \quad (15)$$

où C_i désigne la classe formée au niveau i. La hiérarchie orientée H_A est *significative* au niveau k si sa qualité $q_k(H_A)$ admet un minimum local.

6. Une application

Nous appliquons ici la méthodologie que nous venons de développer à une hiérarchie orientée présentée dans [Gras , Kuntz et Briand., 2003a]**,** portant sur les résultats d'une enquête de l'Association des Professeurs de Mathématiques auprès de 311 professeurs de Terminales. Ceux-ci font choix de 6 objectifs parmi 15 qu'ils assignent à leurs enseignements (ex : A : « Acquisition de connaissances », B : « Préparation à la vie professionnelle ») et d'opinions relatives à dix phrases communément énoncées (ex : OP 1 : « les maths constituent un instrument de sélection excessif ») [Bodin et Gras., 1999]. Les poids des *26* attributs figurent dans TAB 2, compte tenu des pondérations décimales accordées aux variables ordinales (rangs de 6 choix pondérés par 1, 0.8, 0.6, etc. et accords modulés 1,0.5,0 suivant l'accord avec les opinions).

.

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*
105.7	8.8	9.7	140.0	21.8	138.7	19.5	44.8	83.1	108.4	77.6	4.6	90.2

N	*O*	*OP1*	*OP2*	*OP3*	*OP4*	*OP5*	*OP6*	*OP7*	*OP8*	*OP9*	*OPX*	*PER*
66.6	33.2	81.5	147.5	242.5	229.0	190.0	240.0	200.0	165.0	98.0	207.0	254

TAB 2 – Occurrences des attributs de l'enquête sur les professeurs de mathématiques

La hiérarchie orientée obtenue avec le logiciel CHIC [Couturier, 2001] comporte *16* niveaux (figure 1). Le tableau 3 donne les cohésions des *R*-règles correspondantes. Par ex., au niveau 13, la *R*-règle met l'accent sur la relation dérivée des comportements d'ouverture des élèves (I : esprit critique, E : imagination et créativité) vers des situations mathématiques les réalisant :OP8 : exemple et contre-exemple personnels, OP7 : test de réfutation). Cette interprétation globale est difficile par le seul emploi du graphe implicatif qui opère de façon binaire. Ainsi, il y a complémentarité et non redondance entre les deux approches.

Niveaux	*R-règles*	*cohesion*	*Maxima locaux de co*
1	OP8 → OP9	0.998	0.499
2	OP5 → OP4	0.981	
3	N → A	0.955	
4	OP2 → (OP5 → OP4)	0.941	0.821
5	OP9 → OPX	0.92	
6	H → PER	0.92	0.5
7	F → OP3	0.903	
8	(N → A) → OP6	0.865	0.8
9	B → K	0.858	
10	E → (OP8 → OP7)	0.856	0.831
11	G → OP1	0.783	
12	J → (OP9 → OPX)	0.752	
13	I → (E → (OP8 → OP7))	0.707	0.752
14	C → O	0.669	
15	M → (F → OP3)	0.661	0.823
16	L → (J → (OP9 → OPX))	0.404	

TAB 3 – Cohésion des R-règles associées aux niveaux de la hiérarchie de la figure 1 et maxima locaux de l'indice de cohésion-cohérence *co*

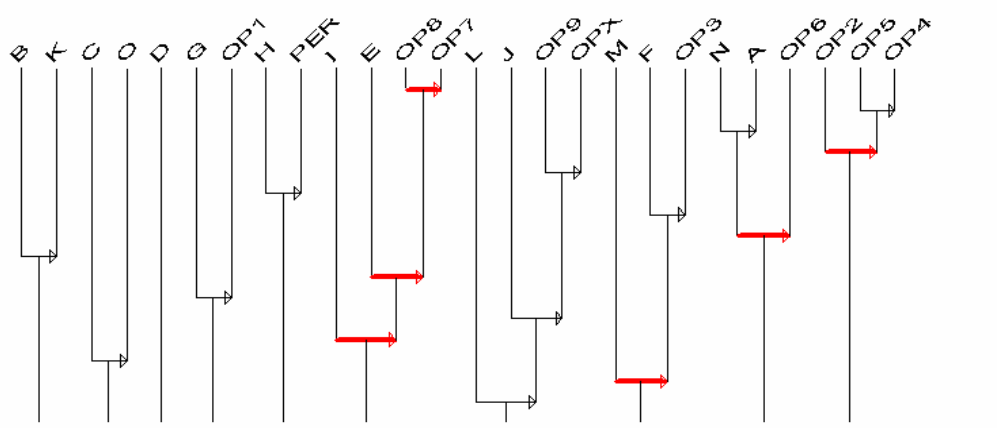

FIG 1 – Hiérarchie orientée obtenue pour l'enquête auprès des professeurs de mathématiques

Le calcul des cohérences à l'aide de l'algorithme implémenté dans CHIC conduit aux probabilités suivantes : $Pr(I_2 > 0) = Pr(I_2 \geq 1) = Pr(I_3 > 1) = 0.5$, $Pr(I_3 > 0) = 0.833$ et $Pr(I_4 > 2) = 0.625$. Une inversion seulement, par rapport aux occurrences, est observée pour les classes des niveaux *12*, *13* et *16*. Les maxima locaux de la cohésion-cohérence sont indiqués dans TAB *3*. On observe également des maxima de l'indice de qualité *q* aux niveaux 1, *4*, *8*, *10*, *13* et *16*. Les niveaux significatifs, indiqués en gras sur FIG 1, à l'exclusion du niveau intéressant *6 (*déclarer la non-pertinence des objectifs, c'est choisir de secondariser le développement de la volonté et la persévérance), avaient déjà été obtenus précédemment avec la méthode globale inspirée des travaux de [Lerman, 1981]. Cependant, ce résultat n'a pas valeur de généralité. Dans la situation expérimentale,la sémantique semble bien respectée dans les deux cas.

7. Conclusion

Dans cet article, nous avons développé une approche pour évaluer la significativité des niveaux d'une hiérarchie orientée et la qualité d'une hiérarchie orientée partielle qui tient compte du préordre défini sur les attributs de chaque *R*-règle constituée à chaque niveau de la hiérarchie. Cette approche ne nécessite pas, contrairement à une approche globale précédemment employée, la détermination d'une préordonnance sur l'ensemble des couples selon le critère de cohésion. De plus, lorsque le nombre *m* d'attributs « classés » devient grands les calculs du nouveau critère peuvent être simplifiés par le recours à une loi normale. Cette approche pourrait être généralisée à la recherche d'une mesure de distorsion entre deux permutations, prolongeant ainsi des travaux de Kendall. Mais, de nouvelles mises à l'épreuve sur des données réelles et, en particulier, des corpus de grande taille tels qu'on les trouve en fouille de données, permettront une comparaison plus robuste sur le plan de l'information restituée au cours des analyses que pourraient en faire des experts des domaines étudiés.

Références

[Agrawal et al., 1993] R. Agrawal, T. Imieliensky et A. Swami. Mining association rules between sets of items in large databases. *Proc. of the ACM SIGMOD'93*, p. 679-696, AAAI Press, 1996.

[Bodin et Bodin., 1999] A. Bodin, R. Gras. Analyse du préquestionnaire enseignants avant EVAPM-Terminales. Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement Public, (425) : 772-786, 1999.

[Couturier, 2001] R. Couturier. Traitement de l'analyse statistique implicative dans CHIC. Actes des Journées sur la Fouille des Données par la Méthode d'Analyse Statistique Implicative, p. 33-50, IUFM Caen, 2001.

[Gras, 1979] R. Gras. 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, 1979.

[Gras et Larher, 1992] R. Gras et A. Larher. L'implication statistique, une nouvelle méthode d'analyse de données. Mathématique, Informatique et Sciences Humaines, 120 : 5-31, 1992.

[Gras, 1997] R. Gras. Nœuds et niveaux significatifs en analyse statistique implicative. Prébublication 97-32, Institut de Recherche de Mathématiques de Rennes, 1997.

[Gras, Kuntz et Briand, 2003a] R. Gras, P. Kuntz et H.Briand. Hiérarchie orientée de règles généralisées en analyse implicative. Extraction des Connaissances et Apprentissage, 17(1) : 145-157, 2003.

[Gras et Kuntz, 2003b] R. Gras et P. Kuntz. Discovering R-rules with an oriented hierarchy., Proc. of the 4th Int. Conf. on Knowledge Discovery and Discrete Mathematics, JIM'03, INRIA, p. 223-229, 2003.

[Gras et Ratsimba-Rajohn, 1996] R. Gras et H. Ratsimba-Rajohn. Analyse non symétrique de données par l'implication statistique. RAIRO-Recherche Opérationnelle, 30(3) : 217-232, 1996.

[Horschka et Klögsen, 1991] P. Horschka et W. Klögsen. A support system for interpreting statistical data. Knowledge Discovery in Databases, p. 325-345, AAAI Press, 1991.

[Kendall, 1938] M.G. Kendall A new measure of rank correlation, Biometrika, XXX p 81-93

[Kendall et Stuart, 1991] M.G. Kendall et A. Stuart. Kendall's advanced theory of statistics. Vol. 2, Edward Arnold, London, 1991.

[Lehn, 2000] R.Lehn. 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, 2000.

[Lent et al., 1997] B. Lent, A.N. Swami, et J. Widow.Clustering association rules. Proc. of the 13th Int. Conf. on Data Engineering, p. 220-231, 1997.

[Lerman, 1981] I.C. Lerman. Classification et analyse ordinale des données. Dunod, Paris

[Siegel, 1956] S. Siegel Nonparametric statistics for the behavioural sciences. New York: McGraw-Hill Book Co., 1956

[Toivonen et al., 1995] H. Toivonen, M. Klementtinen, P. Ronkairen, K. Hätönen et H. Manila. Pruning and grouping of discovered association rules. Workshop notes of the ECML Workshop on Statistics, Machine Learning and Knowledge Discovering in Databases, p. 47-52, 1995.