View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by OpenEdition

Mathématiques et sciences humaines
Mathematics and social sciences

154 | Été 2001
Analyse statistique implicative

Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données

Foundations of the implicative statistical analysis and their extensions for data mining

Régis Gras, Pascale Kuntz and Henri Briand

Electronic version
URL: http://journals.openedition.org/msh/2849
DOI: 10.4000/msh.2849
ISSN: 1950-6821

Publisher
Centre d'analyse et de mathématique sociales de l'EHESS

Printed version
Date of publication: 1 March 2001
ISSN: 0987-6936

Electronic reference
Régis Gras, Pascale Kuntz and Henri Briand, « Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données », *Mathématiques et sciences humaines* [Online], 154 | Été 2001, Online since 10 February 2006, connection on 03 May 2019. URL : http://journals.openedition.org/msh/2849 ; DOI : 10.4000/msh.2849

© École des hautes études en sciences sociales

Math. & Sci. hum., (39e année, n° 154-155, *2001*, p. 9-29)

LES FONDEMENTS DE L'ANALYSE STATISTIQUE IMPLICATIVE ET QUELQUES PROLONGEMENTS POUR LA FOUILLE DE DONNÉES

Régis GRAS[1], Pascale KUNTZ[1], Henri BRIAND[1]

RÉSUMÉ — *L'analyse statistique implicative d'une part permet de déceler les règles pertinentes à partir d'un test d'hypothèse sur des données variées, d'autre part offre, selon une démarche calquée sur la classification hiérarchique classique, une représentation hiérarchique des méta-règles et une analyse des contributions des attributs et individus aux différentes associations. De plus, une nouvelle mesure permet d'intégrer à la fois la notion d'entropie et l'intensité d'implication, adaptée au traitement de données volumineuses telles qu'elles apparaissent en Extraction de Connaissances dans les Données.*

MOTS-CLÉS — Règles, Méta-règles, Quasi-implication, Intensité d'implication, Hiérarchie implicative, Contributions, Entropie.

SUMMARY — Foundations of the implicative statistical analysis and their extensions for data mining

Implicative statistical analysis allows the discovery of relevant rules from a hypothesis test on various data, and offers following classical hierarchical classification, a hierarchical representation of meta-rules and an analysis of the individual and attribute contributions to the different associations. Moreover, a new measure takes into account both the notion of entropy and the intensity of implication, adequate to large data sets studied in data mining.

KEYWORDS — Rules, Meta-rules, Quasi-implication, Intensity of implication, Hierarchy, Contributions, entropy.

1. INTRODUCTION

La recherche de règles imprécises, ou partielles, entre des attributs décrivant une population $\boldsymbol{I}$ d'individus connaît depuis quelques années un regain notable d'intérêt dû, en grande partie, à l'importance croissante des activités en Extraction de Connaissances dans les Données (ECD). Suite notamment aux travaux de Agrawal et coll. [1], ces règles sont devenues un des principaux concepts utilisés pour représenter des connaissances extraites d'une base de données. Intuitivement, il s'agit de découvrir, dans la base, des règles non symétriques pertinentes pour modéliser des relations du type " si

[1] IRIN - École Polytechnique de l'Université de Nantes, La Chantrerie, B.P. 60602, 44306 Nantes cedex□, e-mail : gras@univ-rennes1.fr, Pascale.Kuntz@polytechn.univ-nantes.fr, Henri.Briand@polytechn.univ-nantes.fr

a alors presque b". Contrairement aux approches initiales de l'Analyse combinatoire des données et de la Classification conceptuelle, la condition d'inclusion $\boldsymbol{I}_a \subset \boldsymbol{I}_b$ entre le sous-ensemble d'individus $\boldsymbol{I}_a \subset \boldsymbol{I}$ décrits par a et le sous-ensemble d'individus $\boldsymbol{I}_b \subset \boldsymbol{I}$ décrits par b est ici relaxée et, le traitement des exceptions est alors généralement effectué *a posteriori* par des algorithmes spécifiques (ex. [23]).

En ECD, la probabilité conditionnelle – appelée dans ce contexte " confiance " – est associée au support, le moteur de la définition de l'association le plus souvent utilisé. Le problème étant classique dans d'autres domaines, en particulier l'analyse de questionnaires, d'autres mesures (Loevinger, ...) ont été proposées (ex. [3]). Cependant, à notre connaissance, de nombreux travaux se focalisent sur la recherche d'implications partielles entre les données, souvent binaires, et ne se prolongent ni à l'extraction et à la représentation de méta-règles de type " $a \Rightarrow (b \Rightarrow c)$ ", ni à la recherche des individus ou catégories d'individus responsables des associations.

Dans cet article, nous décrivons une méthodologie, "□analyse statistique implicative ", qui, d'une part permet de déceler les règles pertinentes à partir d'un test d'hypothèse sur des données variées, d'autre part propose, selon une démarche calquée sur la classification hiérarchique classique, une représentation hiérarchique des méta-règles et une analyse des contributions des attributs et individus aux différentes associations.

Inspirés de l'analyse de la vraisemblance du lien sur des données de I. Lerman [19], les premiers travaux ([7], [20]) ont défini une mesure de la relation de type "□⇒□ " basée sur l'invraisemblance de l'apparition, dans les données, du nombre de cas qui l'infirment. Cette mesure vise à quantifier l'" étonnement " d'un expert devant le nombre invraisemblablement petit de contre-exemples eu égard à une indépendance présumée et aux effectifs en jeu. Des prolongements, sous-tendus principalement par des interrogations posées en didactique des mathématiques, ont conduit à la construction du " graphe d'implication " et de la " hiérarchie implicative " ([8], [9], [10], [11]) et différentes généralisations ont été proposées pour des données non binaires ([2], [16], [17]). En parallèle, différents travaux ont montré ces dernières années la pertinence de la démarche en ECD ([6], [13], [15], [25]).

Après un rappel rapide de la mesure d'intensité d'implication et de ses principales propriétés, nous développons les points pré-cités – graphe d'implication, hiérarchie implicative, contributions des individus et des attributs – dans un cadre théorique unifié pour des données binaires. Nous avons privilégié ici une présentation de type axiomatique afin de mettre en avant, d'une part les intuitions trouvant leurs origines dans la manipulation de corpus variés de données réelles qui ont guidé notre démarche, d'autre part les contraintes d'interprétation associées à des propriétés théoriques.

Nous proposons ensuite une nouvelle mesure, intégrant la notion d'entropie, mieux adaptée au traitement de données volumineuses telles qu'elles apparaissent en ECD. Un dernier paragraphe synthétise les extensions proposées pour les variables

modales ordonnées, numériques et les travaux en cours sur le partitionnement d'intervalles.

2. INTENSITÉ D'IMPLICATION – RAPPELS

Considérons un ensemble fini $\boldsymbol{I}$ d'individus décrits par un ensemble fini $\boldsymbol{A}$ d'attributs binaires (de traits, caractères, ...). Dans une situation exemplaire de didactique, il s'agit des performances (réussite/échec) des élèves à des items d'un questionnaire. On note $a(i) = 1$ si l'individu $i \in \boldsymbol{I}$ possède l'attribut $a \in \boldsymbol{A}$ et $a(i) = 0$ sinon.

La règle $a \Rightarrow b$, où a et b sont des attributs de $\boldsymbol{A}$, est logiquement vraie si, tous les individus possédant a possèdent également b c'est-à-dire en termes ensemblistes si l'ensemble $\boldsymbol{I}_a$ des individus $i \in \boldsymbol{I}$ pour lesquels $a(i) = 1$ est contenu dans l'ensemble $\boldsymbol{I}_b$ des individus j pour lesquels $b(j) = 1$.

Cependant, cette inclusion stricte n'est qu'exceptionnellement observée dans la réalité. Il est en effet courant d'observer quelques rares individus possédant a mais pas b sans que ne soit contestée la tendance générale à avoir b quand on a l'attribut a. Relativement au cardinal n de $\boldsymbol{I}$, mais aussi aux cardinaux n_a de $\boldsymbol{I}_a$ et n_b de $\boldsymbol{I}_b$, c'est donc le poids des contre-exemples observés, noté $n_{a \wedge \neg b}$, qu'il faut prendre en compte pour accepter statistiquement de conserver ou non la règle $a \Rightarrow b$. Pour cela, à l'instar de la méthode de vraisemblance du lien de I. Lerman ([18], [19]) nous avons défini une mesure de la tendance implicative $a \Rightarrow b$ à partir de l'invraisemblance de l'apparition dans les données du nombre de cas qui l'infirment ([7], [9]). Et, dans la suite de cet article, la notation $a \Rightarrow b$ désigne sans ambiguïté une “ quasi-implication ”.

L'évaluation de la pertinence de la quasi-implication est basée sur un test d'hypothèse où l'on compare les contre-exemples observés au nombre de contre-exemples attendus sous une hypothèse H_0 d'indépendance. Considérons un tirage aléatoire de deux parties quelconques X et Y de $\boldsymbol{I}$ choisies indépendamment de même cardinaux respectifs que n_a et n_b. Soient $\neg Y$ et $\neg \boldsymbol{I}_b$ les complémentaires respectifs de Y et $\boldsymbol{I}_b$ dans $\boldsymbol{I}$ de même cardinal $n_{\neg b}$. Par conséquent, $\text{card}(X \cap \neg Y)$ est une variable aléatoire dont $n_{a \wedge \neg b}$ est une valeur observée.

DÉFINITION 1. La règle $a \Rightarrow b$ est dite *admissible* au seuil de confiance $1 - \alpha$ si la probabilité que le nombre de contre-exemples dans les observations soit supérieur au nombre de contre-exemples attendus sous l'hypothèse H_0 d'indépendance est faible c'est-à-dire si :

$$\Pr\left(Q(a, b) \leq n_{a \wedge \neg b}\right) \leq \alpha$$

où $Q(a, b)$ est le nombre aléatoire de contre-exemples à l'implication.

La distribution de la variable aléatoire $Q(a, b)$ dépend des hypothèses de tirage [19] : elle peut suivre une loi hypergéométrique, binomiale, ou de Poisson. Dans le cas hypergéométrique $Q(a, b) = Q(b, a)$. Nous préférons une loi non-symétrique car,

comme le notent Zighed et Rakotomalala [25], " l'étude de la rareté des contre-exemples n'est pas simplement le dual de l'étude de l'abondance des exemples ". Nous nous restreignons ici à la loi de Poisson – de paramètre $\lambda = n_a n_{\neg b} / n$ – mais d'autres travaux portent sur la loi binomiale [7] et les résultats expérimentaux sont assez proches. Dans ce cas, $Q(a, b) = \text{card}(X \cap \neg Y)$.

Pour faciliter les calculs, nous considérons dans la suite, dans le cas où $n_{\neg b}$ n'est pas nul, la variable centrée réduite :

$$\frac{\text{card}(X \cap \neg Y) - \lambda}{\sqrt{\lambda}}$$

que nous continuons à noter $Q(a, b)$ pour alléger les notations.

Ainsi, dans les cas légitimant l'approximation (par exemple $\lambda > 10$), la variable aléatoire $Q(a, b)$ peut être approchée asymptotiquement par la loi normale centrée réduite. Et, dans la réalisation expérimentale, la valeur observée de $Q(a, b)$ est alors $(n_{a \wedge \neg b} - \lambda) / \sqrt{\lambda}$, notée $q(a, b)$.

De la définition 1 et de la formule précédente on déduit donc une mesure quantitative de la qualité de la règle $a \Rightarrow b$ ainsi qu'un test statistique pour admettre ou non cette règle.

DÉFINITION 2. On appelle *intensité d'implication* de a sur b la mesure

$$\varphi(a, b) = 1 - \Pr\,(Q(a, b) \leq q(a, b)) \quad \text{pour} \quad n_b \quad n$$

Et, la quasi-implication $a \Rightarrow b$ est *admissible* au niveau de confiance $1 - \alpha$ si et seulement si $\varphi(a, b) \geq 1 - \alpha$.

REMARQUE 1. Lorsque $Q(a, b)$ est approximée par une loi normale alors

$$\varphi(a, b) = \frac{1}{\sqrt{2\pi}} \int_{q_{ab}}^{\infty} e^{-t^2/2}\, dt$$

Lorsque la règle est triviale, comme dans le cas où l'ensemble $\boldsymbol{I}_b$ est très grand ou coïncide avec l'ensemble de tous les individus $\boldsymbol{I}$, alors l'étonnement de constater la rareté du nombre des contre-exemples, en regard du nombre surprenant des instances de l'implication doit être faible.

PROPOSITION 1 [9]. Supposons n_a fixé et $\boldsymbol{I}_a \subset \boldsymbol{I}_b$. Si n_b tend vers n alors l'intensité d'implication $\varphi(a, b)$ tend vers 0.

Pour une comparaison approfondie avec d'autres indices classiques de la littérature, nous renvoyons à [9]. Rappelons tout de même que la valeur de l'intensité d'implication varie avec la dilatation des ensembles $\boldsymbol{I}$, $\boldsymbol{I}_a$ et $\boldsymbol{I}_b$ ce qui n'est pas le cas

de l'indice de Loevinger [21] ni de la probabilité conditionnelle et de toutes ses extensions.

3. GRAPHE D'IMPLICATION

La relation d'admissibilité définie ci-dessus est réflexive et symétrique mais non-transitive. Or, il est utile dans la pratique de mettre en évidence des relations d'ordre partiel entre des sous-ensembles d'attributs. Si $a \Rightarrow b$ et $b \Rightarrow c$, Gras et coll. [9] acceptent la fermeture transitive $a \Rightarrow c$ si $\varphi(a, c) \geq 0.5$, c'est-à-dire si la tendance implicative de a sur c est meilleure que la neutralité.

Pour faciliter l'interprétation et mettre en évidence les relations significatives, les implications avec une intensité d'implication supérieure à 0.5 peuvent être représentées par un graphe orienté G tel que l'ensemble des sommets $\boldsymbol{S}$ est un sous-ensemble de $\boldsymbol{A}$ et tel qu'il existe un arc entre deux sommets a et b si $a \Rightarrow b$ et il n'existe pas c tel que $a \Rightarrow c$ et $c \Rightarrow b$. Ce graphe est sans circuit si les valeurs prises par l'intensité d'implication sont différentes. Par exemple, si $\boldsymbol{S} = (a, b, c, d, e)$ et s'il existe au seuil 0.5 les relations suivantes : $d \Rightarrow c, b, a, e$; $c \Rightarrow a, e$; $b \Rightarrow a, e$; $a \Rightarrow e$ alors les relations sont représentées par le graphe figure 1. Pour ne pas surcharger la représentation, les relations transitives ne sont pas tracées.

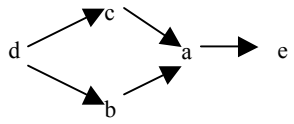

Figure 1. Exemple de graphe d'implication

4. IMPLICATION ENTRE CLASSES

Si le graphe d'implication permet de représenter synthétiquement les relations d'implications entre des couples d'attributs, en revanche il ne permet pas de déceler des implications d'un autre ordre de type $A \Rightarrow B$ où l'un au moins des deux termes est déjà une règle. Pour guider l'intuition, un parallèle peut être établi avec la théorie de la démonstration où une implication $a \Rightarrow (b \Rightarrow c)$ peut décrire une implication entre une propriété a et un théorème $b \Rightarrow c$ précédemment validé.

Il s'agit donc de mettre en évidence les relations d'implications non-symétriques qui existent entre des “ classes ” d'attributs de $\boldsymbol{A}$. Pour rendre compte du mécanisme incrémental de la construction de ces implications d'une part, et de la notion de proximité entre des “ classes ” d'attributs d'autre part, nous avons suivi une démarche similaire à celle de la classification hiérarchique. Cependant, les éléments constitutifs de la hiérarchie ne sont plus ici des parties de l'ensemble des attributs $\boldsymbol{A}$ mais des “□lasses□ d'attributs qui entretiennent entre elles des relations non symétriques.

4.1 CLASSES IMPLICATIVES

Par simplification, nous désignons la règle $a \Rightarrow b$ entre deux attributs quelconques de $\boldsymbol{A}$ par le couple (a, b). Dans la suite, un tel couple est appelé une classe de cardinal 2 et l'ensemble de tous les couples d'éléments distincts engendrés par les attributs de $\boldsymbol{A}$ est noté $\boldsymbol{C}_2$.

Considérons maintenant l'ensemble $\boldsymbol{C}_3$ des classes de cardinal 3 tel que chaque classe est un couple dont l'une des composantes soit un attribut et l'autre composante un élément de $\boldsymbol{C}_2$. Par exemple, si $\boldsymbol{A} = \{a, b, c\}$ alors $\boldsymbol{C}_2 = \{(a, b), (b, a), (a, c), (c, a), (b, c), (c, b)\}$ et $\boldsymbol{C}_3 = \{(a, (b, c)), ((a, b), c), (a, (c, b)), ((a, c), b), (b, (a, c)), ((b, a), c), (b, (c, a)), ((b, c), a), (c, (a, b)), ((c, a), b), (c, (b, a)), ((c, b), a)\}$. Par exemple, la classe $(a, (b, c))$ désigne la règle $a \Rightarrow (b \Rightarrow c)$.

Pour tenir compte de relations quasi-implicatives plus globales, nous pouvons généraliser cette définition.

DÉFINITION 3. Posons $\boldsymbol{C}_1 = \boldsymbol{A}$. L'ensemble $\boldsymbol{C}_i$, $2 \leq i \leq n$ où n est le cardinal de $\boldsymbol{A}$, des classes de cardinal i est composé des couples dont les composantes distinctes appartiennent à $\boldsymbol{C}_j$, $1 \leq j < i$ et tels que le nombre d'attributs total des deux composantes est égal à i. On note $\boldsymbol{C} = \boldsymbol{C}_2 \cup \boldsymbol{C}_3 \cup ... \cup \boldsymbol{C}_n$ l'ensemble de toutes les classes engendrées à partir de $\boldsymbol{A}$ et $C = (A, B)$ une classe quelconque de $\boldsymbol{C}$.

On appelle ordre implicatif $>_I$ sur l'ensemble des attributs de C l'ordre défini par la lecture de gauche à droite de la relation d'implication.

Par exemple, si $\boldsymbol{A} = \{a, b, c, d\}$, les classes $((a, b), (c, d))$ et $(a, (b, (c, d)))$ sont des éléments de $\boldsymbol{C}_4$ désignant respectivement les règles $(a \Rightarrow b) \Rightarrow (c \Rightarrow d)$ et $a \Rightarrow (b \Rightarrow (c \Rightarrow d))$. L'ordre implicatif est le même pour ces deux classes : $a >_I b >_I c >_I d$.

Notre objectif consiste alors à mettre en évidence les classes de $\boldsymbol{C}$ dont l'implication entre les composantes est significative au sens d'un certain critère de "cohésion ". En prenant une métaphore hydraulique, on souhaite que le " flux " implicatif d'une classe C sur une classe C' soit nourri d'un " flux " interne à C et alimente un " flux " interne à C'. Cette dynamique interne orientée ne peut être prise en compte par un indice de similarité classique symétrique et nous devons donc définir une nouvelle mesure.

4.2. COHÉSION ET IMPLICATION ENTRE CLASSES

Intuitivement, la cohésion doit être un indicateur d'un certain ordre implicatif au sein d'un sous-ensemble d'attributs. Elle s'oppose en cela au " désordre " dont rend compte l'entropie d'une expérience aléatoire. Ainsi, la définition choisie relie la notion d'implication à celle de l'entropie. Nous commençons tout d'abord par définir la

cohésion d'une classe (a, b) de cardinal 2 puis étendons la définition à des classes quelconques de $\boldsymbol{C}$.

Soit X la variable aléatoire indicatrice de l'évènement $(Q(a, b) \geq q(a, b))$. La loi de X est définie par $\Pr(X = 1) = \varphi(a, b) = p$ et $\Pr(X = 0) = 1 - \varphi(a, b) = 1 - p$. Et l'entropie de cette expérience est $I(X) = - p \log_2 p - (1 - p) \log_2(1 - p)$.

Si $\varphi(a, b) = 1$ alors l'entropie est nulle (en convenant que $0\log0 = 0$) et si $\varphi(a, b) = 0.5$ alors $I(X)$ est maximale et vaut 1. La valeur $\varphi(a, b) = 0.5$ coïncide avec la situation où $n_{a \wedge \neg b} = n_a . n_{\neg b} / n$, ce qui signifie que l'observation de $n_{a \wedge \neg b}$ est égale à la valeur moyenne attendue, c'est-à-dire sans surprise. Dans ce cas et lorsque $\varphi(a, b) < 0.5$, il nous paraît naturel d'annuler la cohésion, la liaison implicative ayant perdu toute sa signification.

DÉFINITION 4. L'*indice de cohésion* d'une classe (a, b) de $\boldsymbol{C}_2$ est définie par

$c(a, b) = (1 - (- p \log_2 p - (1 - p) \log_2(1 - p))^2)^{1/2}$ si $p = \varphi(a, b) > 0.5$

$c(a, b) = 0$ si $p = \varphi(a, b) \leq 0.5$

Le carré de l'entropie est choisi pour renforcer le contraste entre les valeurs sur $[0, 1]$, et la racine carrée du complément à 1 permet de mesurer la cohésion sur une même échelle que l'entropie.

Remarquons que si $n_a < n_b$ alors $\varphi(a, b) > \varphi(b, a)$ et $c(a, b) > c(b, a)$. Dans ce cas, lors de la construction des classes significatives au sens de la cohésion nous conservons la classe (a, b), c'est-à-dire la quasi-implication $a \Rightarrow b$ au détriment de (b, a).

La généralisation de cette définition à des classes de cardinal supérieur est guidée par une exigence : l'indice de cohésion d'une classe doit contenir l'information révélée par les relations quasi-implicatives entre tous les attributs de la classe, tout en respectant l'ordre des quasi-implications entre ces différents attributs. Par exemple, considérons la classe $(a, (b, c))$ de cardinal 3. La cohésion de la classe doit tenir compte des cohésions des classes (a, b), (a, c) et (b, c). Une moyenne entre ces valeurs permet de satisfaire à cette exigence.

Deux propriétés nous conduisent à choisir la moyenne géométrique. Elle permet d'obtenir une cohésion de classe nulle dès que la cohésion est nulle pour un couple d'attributs c'est-à-dire dès qu'une implication mutuelle est faible ou sans surprise. De plus, elle conduit à une valeur proche de 1 lorsque les cohésions des couples sont assez fortes.

DÉFINITION 5. Soit C une classe quelconque de $\boldsymbol{C}$ avec comme attributs dans l'ordre implicatif défini par la classe $c_1, .. c_k$.

L'indice de cohésion de la classe C est défini par :

$$c(C) = \left(\prod_{\substack{i=1,k \\ j=2,k\ ;\ j>i}} c(c_i, c_j) \right)^{\frac{2}{k(k-1)}}$$

À partir de cette définition, on peut généraliser la notion d'implication entre variables à celle d'implication entre classes. Pour des raisons d'interprétation et pour être en phase avec les idées intuitives présentées au début de ce paragraphe, nous souhaitons que l'intensité entre deux classes C et C' intègre trois informations : les cohésions respectives des classes, une intensité d'implication extrémale des éléments d'une classe sur les éléments de l'autre et les cardinaux respectifs des deux classes. Plus précisément, nous souhaitons que cet indice satisfasse trois propriétés :

- il doit croître avec les cohésions de chaque classe et s'annuler lorsque l'une des cohésions est nulle ;
- il doit croître avec la liaison extrémale (minimale si l'on vise un degré d'exigence élevé, maximale si l'on se contente d'une souplesse réaliste) ;
- il doit décroître avec les cardinaux des classes.

DÉFINITION 6. Soit C et C' deux classes quelconques de $\boldsymbol{C}$ où les attributs de C dans l'ordre implicatif défini par la classe sont $c_1, ..c_k$ et ceux de C' sont $c'_1, ...c'_h$.

L'indice d'implication généralisé $\varphi_G(C, C')$ de la classe C vers la classe C' est défini par :

$$\varphi_G(C,C') = \left(\max_{i=1,k\ ;\ j=1,h} \varphi(c_i, c'_j) \right)^{hk} . \left(c(C) \ . \ c(C')\right)^{1/2}$$

Remarquons que, si les deux cohésions sont multipliées par une constante alors il en est de même pour $\varphi_G(C, C')$. L'exposant hk permet de diminuer l'effet des classes à grands effectifs d'attributs.

4.3. HIÉRARCHIE IMPLICATIVE

Partant d'un ensemble $\boldsymbol{I}$ d'individus décrit par un ensemble d'attributs $\boldsymbol{A}$, les classes significatives sont construites par une méthode ascendante hiérarchique directement inspirée des méthodes classiques de classification hiérarchique. Nous conservons par analogie le terme “ hiérarchie ” bien qu'ici les éléments ne soient pas des parties d'un ensemble mais des classes définies au paragraphe 4.2 et que la relation entre les éléments ne soit plus une relation d'inclusion mais une relation de quasi-implication.

Les classes sont construites par agrégations successives basées sur l'indice de cohésion. Ainsi, au premier niveau se réunissent en une classe deux attributs dont la cohésion est maximale parmi toutes les cohésions des autres classes de cardinal 2. Au

niveau suivant, se forme soit une classe de cardinal 2 – celle de cohésion maximale parmi $\boldsymbol{C}_2$ privé de la classe formée au niveau 1 – soit une classe de cardinal 3 dont l'une des composantes est la classe formée au niveau 1 et l'autre un attribut et ainsi de suite.

Plus précisément, décrivons la construction de la classe de niveau $h > 2$. Notons $\boldsymbol{H}_{h-1}=\{C_1, ..., C_{h-1}\}$ l'ensemble des classes construites aux niveaux précédents et k le cardinal maximal de ces classes.

Nous comparons les cohésions suivantes qui correspondent aux classes de cardinal 2, puis de cardinal 3 jusqu'à $2k$ dont au moins une des composantes est une classe C_i construite précédemment :

- pour les classes de cardinal 2 : $c(C_i)$ pour $C_i \in \boldsymbol{C}_2 - \boldsymbol{H}_{h-1} \cap \boldsymbol{C}_2$
- pour les classes de cardinal 3 : $c(a, C_i)$ et $c(C_i, a)$ pour $C_i \in \boldsymbol{H}_{h-1} \cap \boldsymbol{C}_2$ et a attribut de $\boldsymbol{A}$ absent dans C_i
- pour les classes de cardinal 4 : $c(C_i, C_j)$ avec C_i et $C_j \in \boldsymbol{H}_{h-1} \cap \boldsymbol{C}_2$; $c(a, C_i)$ et $c(C_i,$□$)$ pour $C_i \in \boldsymbol{H}_{h-1} \cap \boldsymbol{C}_3$ et a attribut de $\boldsymbol{A}$ absent dans C_i
- pour les classes de cardinal $k' \leq 2k$□ $c(a, C_i)$ et $c(C_i, a)$ pour $C_i \in \boldsymbol{H}_{h-1} \cap \boldsymbol{C}_{k'-1}$ et a attribut de $\boldsymbol{A}$ absent dans C_i ; $c(C_i, C_j)$ où C_i et $C_j \in \boldsymbol{H}_{h-1}$ et sont telles que le cardinal de C_i plus le cardinal de C_j vaut k'.

La classe retenue au niveau h est celle dont la cohésion est maximale et non nulle parmi l'ensemble des classes décrites ci-dessus. En cas d'ex æquo, si deux classes (C_i, C_j) et (C'_i, C'_j) admettent la même cohésion au niveau h, on retient celle qui réalise une liaison implicative maximale entre les attributs respectifs des deux dernières composantes (C_i et C_j ou C'_i et C'_j) déjà constituées au niveau h - 1 antérieur.

L'algorithme s'arrête, contrairement aux hiérarchies classiques construites avec des similarités, dès que toute extension des classes précédemment construites conduit à une cohésion nulle. Nous notons $\boldsymbol{C}^H$ l'ensemble des classes de la hiérarchie H.

Par exemple, sur la hiérarchie implicative de la figure 2, les classes construites successivement sont (e, d), (b, c) et $(a, (e, d))$□ elles correspondent aux règles $e \Rightarrow d$, $b \Rightarrow c$ et $a \Rightarrow (e \Rightarrow d)$. Au deuxième niveau ($h = 2$), la formation de (b, c) résulte du fait que cette classe a la plus grande cohésion parmi les éléments de l'ensemble {$(a,$□$)$, $((e, d), a)$, (a, b), (b, a), $(b, (e, d))$, $(c, (e, d))$, $((e, d), b)$, $((e, d), c)$, (c, b), $(b,$□$)$}. De même au troisième niveau, la classe (a, (e, d)) a la plus grande cohésion parmi les éléments de {$(a, (b, c))$, $((b, c), a)$, $((e, d), (b, c))$, $((b, c), (e, d))$}. Et, la construction de la hiérarchie implicative s'arrête au troisième niveau si $c((a,$□$)$, $(b,$□$)) = c((b, c), (a, (e, d))) = 0$.

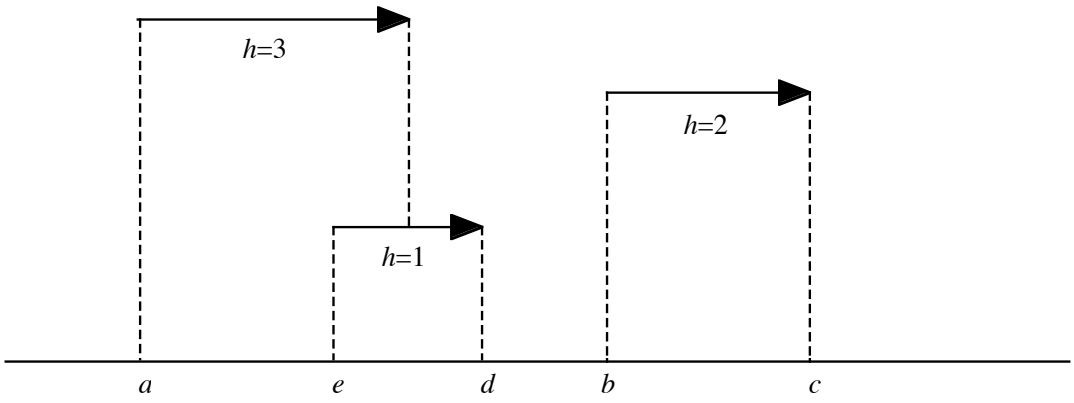

Figure 2. Exemple de hiérarchie implicative

5. DÉTERMINATION DES NIVEAUX SIGNIFICATIFS

Étant donnée la multiplicité des niveaux de la hiérarchie, il est indispensable de dégager ceux qui sont le plus pertinent par rapport à l'intention classificatrice de l'utilisateur et eu égard aux critères choisis. Nous adoptons un critère similaire à celui décrit par I.□erman [19] qui est défini sur une préordonnance.

On peut définir sur l'ensemble $\boldsymbol{C}_2 = \boldsymbol{A} \times \boldsymbol{A}$ - {(*a, a*), (*b, b*), ...} des classes de cardinal 2 une relation de préordre total Ω basée sur la mesure de cohésion : (*a, b*) < (*c,*□) ⇔ *c*(*a,*□) < *c*(*c,*□). Soit $G(\Omega)$ son graphe : $G(\Omega) = \{(C_i, C_j) \in \boldsymbol{C}_2 \, ; \, C_i < C_j\}$.

À un niveau quelconque h de la hiérarchie, certains couples d'attributs de $\boldsymbol{C}_2$ ont été réunis dans une classe au cours des différentes étapes de la construction et d'autres non. Notons $\boldsymbol{S}_h$ l'ensemble des couples séparés au niveau h et $\boldsymbol{R}_h$ l'ensemble des couples réunis. Par définition, $\boldsymbol{S}_h \cup \boldsymbol{R}_h = \boldsymbol{C}_2$.

L'ensemble $G(\Omega) \cap (\boldsymbol{S}_h \times \boldsymbol{R}_h)$ est constitué des couples de couples qui respectent au niveau h le préordre Ω. Par exemple, si pour les attributs a, b, e, f on a $c(e, f) < c(a, b)$ et si au niveau h les attributs k, e et f sont séparés pendant que les attributs a et b sont réunis dans une classe, alors le couple $((e, f), (a, b))$ appartient à $G(\Omega) \cap (\boldsymbol{S}_h \times \boldsymbol{R}_h)$.

De façon similaire à celle adoptée dans le paragraphe 2, nous cherchons à mesurer l'adéquation entre $G(\Omega)$ et $(\boldsymbol{S}_h \times \boldsymbol{R}_h)$. Associons donc au cardinal de $G(\Omega) \cap (\boldsymbol{S}_h \times \boldsymbol{R}_h)$ la variable aléatoire $G(\Omega^*) \cap (\boldsymbol{S}_h \times \boldsymbol{R}_h)$ où Ω^* est une préordonnance aléatoire dans l'ensemble de toutes les préordonnances de même cardinal que Ω muni d'une probabilité uniforme.

Nous sommes ici dans des conditions d'application du théorème de Wald et Wolfowitz [24] similaires à celles décrites par I. Lerman [19] et pouvons en déduire

que $G(\Omega^*) \cap (\boldsymbol{S}_h \times \boldsymbol{R}_h)$ a pour espérance $\mu = 1/2 \text{ card}(\boldsymbol{S}_h \times \boldsymbol{R}_h)$ et pour variance $\sigma^2 = \text{card}(\boldsymbol{S}_h \times \boldsymbol{R}_h)$. (card $G(\Omega)$ +1) /12.

L'adéquation observée entre $G(\Omega)$ et $(\boldsymbol{S}_h \times \boldsymbol{R}_h)$ au niveau h peut donc être mesurée par l'indice de similarité $s(\Omega,h)$ centré réduit suivant :

$$s(\Omega,h) = \frac{\text{card}\big((G(\Omega) \cap (S_h \times R_h)\big) - \mu}{\sigma}$$

DÉFINITION 7. On appelle *niveau significatif* de la hiérarchie H tout niveau correspondant à un maximum local (meilleur que le niveau précédent et le niveau suivant) de l'indice $s(\Omega,h)$.

6. CONTRIBUTIONS DES INDIVIDUS ET DES ATTRIBUTS

Afin de comprendre la structure de la hiérarchie implicative, il est nécessaire de connaître la contribution des attributs et les individus – ou groupes d'individus – à la constitution des classes.

6.1. CONTRIBUTION DES ATTRIBUTS

Soit $C \in \boldsymbol{C}^H$ la classe formée au niveau h de la hiérarchie H. Par construction, cette classe est l'agrégation de deux classes $C' \in \boldsymbol{C}^H$ et $C'' \in \boldsymbol{C}^H$ non agrégées au niveau $h-1$.

On appelle *couple générique* de niveau h de C le couple d'attributs (a, b) tel que $\varphi(a, b) > \varphi(u, v)$ pour tout attribut u de C' et attribut v de C'', et $\varphi_h(C) = \varphi(a,b)$ *l'intensité générique* de niveau h de C. Ce couple caractérise l'effet implicatif le plus apparent dans C.

Mais les classes C' et C'' dont l'agrégation conduit à C peuvent être elles-mêmes issues de l'agrégation de classes de niveau inférieur. Ainsi, à chaque niveau $k<h$ de la hiérarchie, on peut déterminer, comme au niveau h, un couple générique porteur de la liaison maximale au sein de la classe constituée à ce niveau et définir l'intensité générique $\varphi_k(C)$ de niveau k.

Le vecteur $(\varphi_1(C), ..., \varphi_h(C)) \in [0,1]^h$ est appelé *vecteur puissance implicative* de la classe C formée au niveau h.

L'introduction de ces vecteurs permet de construire un espace de représentation muni d'une métrique où les individus – comme il est fait de façon duale en analyse des correspondances – peuvent être projetés.

6.2. CONTRIBUTION D'UN INDIVIDU

Nous cherchons tout d'abord à définir une dissimilarité sur $\boldsymbol{I} \times \boldsymbol{C}^H$ qui quantifie la part de responsabilité d'un individu quelconque dans la construction d'une classe.

La construction d'une classe C au niveau h se déduisant d'agrégations successives, commençons par rendre compte du respect ou non d'un individu $i \in \boldsymbol{I}$ de la quasi-implication du couple générique (a, b) de C de niveau h.

En fonction de la présence ou non des attributs a et b pour i on peut caractériser la contribution de i au couple (a, b) par la fonction suivante :

$\varphi_{i,h}(C) = 1$ si $a(i) = 1$ ou 0 et $b(i) = 1$
$\varphi_{i,h}(C) = 0$ si $a(i) = 1$ et $b(\mathrm{i}) = 0$
$\varphi_{i,h}(\mathrm{C}) = \mathrm{p} \in]0,1[$ si $a(i) = b(i) = 0$

Le plus souvent nous choisissons la valeur " neutre " p = 0.5.

Ainsi, on peut associer à un individu i, un vecteur à h composantes $\varphi_{i,1}(C),...,$ $\varphi_{\mathrm{i,h}}(C)$ qui caractérisent les contributions de i aux h couples génériques de la classe C. Un individu dont toutes les composantes valent 1 est appelé individu idéal théorique et noté i_t.

Avec cette représentation, il suffit pour caractériser la part de responsabilité d'un individu dans la construction d'une classe de choisir une métrique sur $[0,1]^h$. Nous choisissons la distance du χ^2 entre les distributions $(1 - \varphi_k(C))$ et $(1 - \varphi_{i,k}(C))$ car elle permet, contrairement à la métrique euclidienne habituelle, de relativiser à φ_k l'écart $\varphi_k - \varphi_{i,k}$ et donc de majorer, par exemple, l'effet de cet écart sur la distance lorsque l'intensité φ_k est elle-même très forte.

DÉFINITION 8. On appelle *distance implicative* d'un individu $i \in \boldsymbol{I}$ à la classe $C \in \boldsymbol{C}^H$ le nombre

$$d(i,C) = \left(\frac{1}{h} \sum_{k=1}^{h} \frac{(\varphi_k - \varphi_{i,k})^2}{1 - \varphi_k} \right)^{1/2}$$

S'il existe k tel que $\varphi_k(C) = 0$ nous attribuons la valeur nulle au quotient $(\varphi_k - \varphi_{i,k})^2 / (1 - \varphi_i)$. Cette convention est cohérente avec l'interprétation de $\varphi_k(C)$ puisque, dans ce cas, l'implication générique est maximale et significative d'une excellente liaison implicative entre ses deux termes, vérifiée en général par tous les individus i de $\boldsymbol{I}$ $(\varphi_{i,k}(C) = 1)$.

Nous déduisons de cette distance une mesure de *contribution* $\gamma(i, C)$ d'un individu i à une classe C qui rend compte de la distorsion de i par rapport à l'individu idéal théorique i_t :

$$\gamma(i,C) = \frac{d(i_t,C)}{d(i,C)}$$

Lorsqu'il existe des individus i à distance nulle de C, on considère que la contribution est fixée à une valeur maximale. Lorsque la contribution est supérieure à 1 pour un individu, on peut en déduire que son comportement est plus adéquat à la tendance générale, dans le cadre notionnel des couples génériques, que les individus idéaux théoriques.

On pourrait proposer une autre mesure de la contribution qui rendrait compte de la responsabilité d'un individu de l'existence d'une classe $C \in \boldsymbol{C}^H$:

$$\gamma'(i,C) = 1 - \frac{1}{h} \sum_{k=1}^{h} (1 - \varphi_{i,k})^2$$

Cette contribution a pour valeur 1 lorsque (C) vaut 1 pour tous les couples génériques de C.

6.3. GROUPE D'INDIVIDUS OPTIMAL D'UNE CLASSE

Partant de la définition précédente de la contribution d'un individu à une classe, nous pouvons définir la contribution $\gamma(G, C)$ d'un groupe d'individus $G \subset \boldsymbol{I}$ à la classe C par le barycentre des valeurs de contribution des individus de G à C :

$$\gamma(G,C) = \frac{1}{\text{card}(G)} \sum_{i \in G} \gamma(i,C)$$

Dans la pratique, il est nécessaire de disposer d'un outil qui permette de vérifier rapidement si une catégorie d'individus donnée est statistiquement ou non déterminante dans la constitution d'une classe implicative.

L'idée sous-jacente à la construction de cet outil consiste à rechercher dans l'ensemble des individus $\boldsymbol{I}$, une partition en deux groupes I_1 et I_2 s'opposant de façon optimale quant à leur contribution à C par rapport à l'ensemble de toutes les partitions de $\boldsymbol{I}$ en deux groupes. La discrimination entre les contributions $\gamma(I_1, C)$ et $\gamma(I_2, C)$ des deux groupes I_1 et I_2 est fonction de la dispersion entre les deux barycentres qu'elles représentent et peut donc être, dans ce cas, mesurée par une variance inter-classes.

Notons $n(I_1)$ (resp. $n(I_2)$) l'effectif de I_1 (resp. I_2). Le barycentre $\bar{\gamma}$ des contributions $\gamma(I_1, C)$ et $\gamma(I_2, C)$ est ici $\bar{\gamma} = (n(I_1)\,\gamma(I_1, C) + n(I_2)\,\gamma(I_2, C)) / n$. Par définition des contributions, $\bar{\gamma}$ est également le barycentre de l'ensemble des contributions des individus de $\boldsymbol{I}$. La variance inter-classes V_E ou variance relative aux deux barycentres de I_1 et I_2 est donc□

$$V_E = \frac{n(I_1)}{n}\left(\gamma(I_1,C) - \bar{\gamma}\right)^2 + \frac{n(I_2)}{n}\left(\gamma(I_2,C) - \bar{\gamma}\right)^2$$

DÉFINITION 9. Un groupe d'individus optimal $G^*(C) \subset \boldsymbol{I}$ d'une classe $C \in \boldsymbol{C}^H$ est un sous-ensemble de $\boldsymbol{I}$ qui accorde à C une contribution plus grande que son

complémentaire dans I et qui forme avec celui-ci une partition maximisant la variance inter-classes V_E sur l'ensemble des bi-partitions de $\boldsymbol{I}$.

Nous renvoyons à [10] pour la preuve de l'existence de ce groupe optimal et la description d'un algorithme de calcul.

Il est alors intéressant de connaître, parmi des groupes d'individus donnés *a priori* ou définis par un attribut supplémentaire, quel est celui qui contribue le plus fortement au groupe d'individus optimal $G^*(C)$.

Pour ce faire, nous reprenons une démarche semblable à celle décrite dans le paragraphe 2. Il s'agit de mesurer l'étonnement relatif au constat d'une certaine proportion d'individus d'un groupe eu égard à ce qu'il était probable d'attendre.

Soit $\{I_i\}_i$ une partition *a priori* de $\boldsymbol{I}$. Pour chaque classe I_i de cette partition, considérons X_i une partie aléatoire de $\boldsymbol{I}$ de même cardinal que I_i et Z_i la variable aléatoire $\text{card}(X_i \cap G^*(C))$. La variable Z_i suit une loi binomiale de paramètre $n(I_i)$ et $p = \text{card}G^*(C) / n$ où p est la probabilité pour qu'un individu quelconque de I appartienne au groupe optimal $G^*(C)$. Si un groupe admet dans le groupe optimal une représentation plus grande en effectifs que d'une part, celle qui était prévisible selon la loi binomiale ci-dessus et d'autre part, d'autres groupes, nous considérons que ce groupe est le plus typique de C.

DÉFINITION 10. On appelle groupe d'individus le plus typique de la classe C, le sous-ensemble $I_i \subset \boldsymbol{I}$ qui minimise la probabilité $\Pr(Z_i > \text{card}(X_i \cap G^*(C)))$ sur l'ensemble des sous-ensembles de $\boldsymbol{I}$.

7. DONNÉES DE GRANDE TAILLE - VERSION ENTROPIQUE

À l'origine, l'intensité d'implication a été mise en œuvre dans le cadre de la didactique des mathématiques où il s'agissait d'organiser des problèmes en fonction des difficultés ressenties par les élèves. Du fait de la complexité des expérimentations, les corpus traités étaient de taille relativement restreinte. L'extension récente de la méthode à des corpus de grandes tailles, stimulée en particulier par le nouveau champ de recherche de l'ECD, soulève deux problèmes.

Lorsque la taille de $\boldsymbol{I}$ croît, l'intensité d'implication $\varphi(a, b)$ a tendance à ne plus être suffisamment discriminante car ses valeurs peuvent être souvent voisines de 1 alors que l'inclusion $\boldsymbol{I}_a \subset \boldsymbol{I}_b$, dont elle cherche à quantifier la qualité, est loin d'être parfaite. De plus, le modèle du paragraphe 2 retient essentiellement la mesure de la pertinence de la règle $a \Rightarrow b$. Or, la prise en compte de la contraposée $\neg b \Rightarrow \neg a$ permettrait de renforcer l'affirmation de la relation implicative de a sur b. Elle pourrait également contribuer à répondre au premier problème puisque si $\boldsymbol{I}_a$ et $\boldsymbol{I}_b$ sont petits relativement à $\boldsymbol{I}$ leurs complémentaires seront grands et réciproquement.

Pour pallier ces insuffisances, nous proposons de tenir compte de la qualité de l'information fournie par la faiblesse relative des instances qui contredisent la règle et sa contraposée. Plus précisément, il s'agit de moduler la valeur de l'“ étonnement ” en fonction d'une part, du déséquilibre entre $n_{a \wedge b}$ et $n_{a \wedge \neg b}$ associé à $a \Rightarrow b$ et d'autre part, du déséquilibre entre $n_{a \wedge \neg b}$ et $n_{\neg a \wedge \neg b}$ associé à $\neg b \Rightarrow \neg a$. L'étonnement doit être atténué (resp. confirmé) lorsque le nombre $n_{a \wedge \neg b}$ de contre-exemples aux deux règles est élevé (resp. faible) eu égard aux effectifs des observations n_a et $n_{\neg b}$.

Pour mesurer ces déséquilibres de façon non linéaire, un indice classique est l'entropie conditionnelle de Shannon [22]. L'entropie conditionnelle $H_{b/a}$ relative aux cas (a et b) et (a et $\neg b$) lorsque a est vérifié est définie par :

$$H_{b/a} = -\frac{n_{a \wedge b}}{n_a} \log_2 \frac{n_{a \wedge b}}{n_a} - \frac{n_{a \wedge \neg b}}{n_a} \log_2 \frac{n_{a \wedge \neg b}}{n_a}$$

De même, l'entropie conditionnelle $H_{\neg a/\neg b}$ relative aux cas ($\neg a$ et $\neg b$) et (a et $\neg b$) lorsque $\neg b$ est vérifié est définie par :

$$H_{\neg a/\neg b} = -\frac{n_{a \wedge \neg b}}{n_{\neg b}} \log_2 \frac{n_{a \wedge \neg b}}{n_{\neg b}} - \frac{n_{\neg a \wedge \neg b}}{n_{\neg b}} \log_2 \frac{n_{\neg a \wedge \neg b}}{n_{\neg b}}$$

D'une façon générale, ces entropies devraient être simultanément petites si l'on souhaite disposer d'un bon critère d'inclusion de $\boldsymbol{I}_a$ dans $\boldsymbol{I}_b$. En effet, les entropies représentent l'incertitude moyenne des expériences qui consistent à observer si b est réalisé (resp. si non a est réalisé) lorsque l'on a observé a (resp. non b). Le complément à 1 de cette incertitude représente donc l'information moyenne recueillie par la réalisation de ces expériences. Plus cette information est importante, plus forte est la garantie de la qualité de l'implication et de sa contraposée. Cependant, l'impact des déséquilibres doit être ajusté en fonction des différentes situations cardinales. Nous proposons ici une nouvelle définition plus simple formellement qu'une autre récemment publiée [12].

Pour que le modèle ait la signification attendue, il doit satisfaire les contraintes suivantes :

- il doit intégrer les valeurs de l'entropie et, pour les contraster, par exemple, intégrer ces valeurs au carré ;
- comme ce carré varie de 0 à 1, afin de dénoter le déséquilibre et donc l'inclusion, afin de s'opposer à l'entropie, la valeur retenue est le complément à 1 de son carré tant que le nombre de contre-exemples est inférieur à la moitié des observations de a (resp. de $\neg b$). Au-delà de ces valeurs, les implications n'ayant plus de sens inclusif, on affecte au critère la valeur 0 ;
- afin de prendre en compte les deux informations propres à $a \Rightarrow b$ et $\neg b \Rightarrow \neg a$, le produit rend compte de la qualité simultanée des valeurs retenues. Le produit s'annule dès que l'un de ses termes s'annule, c’est-à-dire dès que cette qualité s'efface.

- enfin, le produit ayant une dimension 4 par rapport à l'entropie, sa racine quatrième est de la même dimension.

Posons $\alpha = n_a / n$ (resp. $\beta = n_{\neg b} / n$) la fréquence de a (resp. non b) sur la population. Selon les contraintes précédentes, les deux " ajustements " de l'entropie significatifs des qualités respectives de l'implication et de sa contraposée peuvent donc s'écrire en fonction de la fréquence $t = n_{a \wedge \neg b} / n$ de contre-exemples de la façon suivante :

$$h_1(t) = H(b / a) = -\left(1 - \frac{t}{\alpha}\right) \log_2 \left(1 - \frac{t}{\alpha}\right) - \frac{t}{\alpha} \log_2 \left(\frac{t}{\alpha}\right) \quad \text{si} \quad t \in \left[0, \frac{\alpha}{2}\right]$$

$$h_1(t) = 0 \quad \text{si} \quad t \in \left[\frac{\alpha}{2}, \alpha\right]$$

$$h_2(t) = H(\neg a / \neg b) = -\left(1 - \frac{t}{\beta}\right) \log_2 \left(1 - \frac{t}{\beta}\right) - \frac{t}{\beta} \log_2 \left(\frac{t}{\beta}\right) \quad \text{si} \quad t \in \left[0, \frac{\beta}{2}\right]$$

$$h_2(t) = 0 \quad \text{si} \quad t \in \left[\frac{\beta}{2}, \beta\right]$$

DÉFINITION 11. L' indice d'inclusion de $\boldsymbol{I}_a$, support de a, dans $\boldsymbol{I}_b$, support de b, est le nombre qui intègre l'information délivrée par la réalisation d'un faible nombre de contre-exemples, d'une part à la règle $a \Rightarrow b$ et, d'autre part, à la règle $\neg b \Rightarrow \neg a$.

$$\tau(a, b) = \left(1 - h_1^2(t)\right) . \left(1 - h_2^2(t)\right)^{1/4}$$

L'*intensité entropique* de la règle $a \Rightarrow b$ est définie par $\psi(a, b) = (\varphi(a, b).\ \iota(a, b))^{1/2}$ où $\varphi(a, b)$ est l'intensité d'implication et $\tau(a, b)$ l'indice d'inclusion.

Nous voyons sur la figure 3 qui représente cette nouvelle mesure – pour n_a et $n_{\neg b}$ fixés – que les propriétés attendues sont bien respectées :

- " réaction " lente aux premiers contre-exemples (résistance au bruit) ;
- " accélération " du rejet de l'inclusion au voisinage de l'équilibre ;
- rejet de plus en plus accentué au-delà de l'équilibre, ce que n'assurait pas l'intensité d'implication.

Les deux exemples suivants (tableaux 1.a et 1.b) permettent de comparer la nouvelle mesure $\Psi(a, b)$ avec l'intensité d'implication $\varphi(a, b)$ et la probabilité conditionnelle $\Pr(b / a)$.

Pour les données décrites dans le tableau 1.a, l'intensité d'implication vaut 0.9999. Les valeurs entropiques sont $h_1 = h_2 = 0$, par conséquence l'indice d'inclusion s'annule. Ainsi, $\Psi(a, b) = 0$ alors que $\Pr(b / a) = 0.3333$. Les fonctions entropiques " modèrent " l'intensité d'implication dans le cas où justement l'inclusion est médiocre.

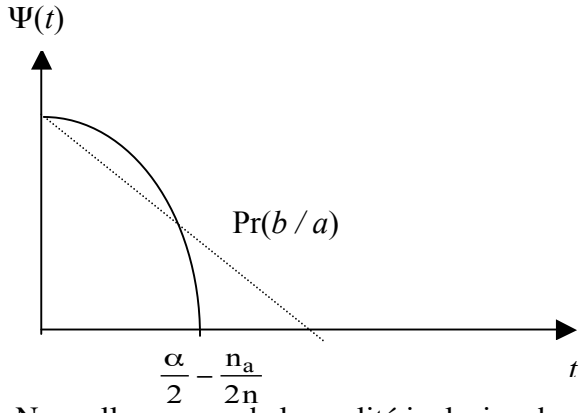

Figure 3. Nouvelle mesure de la qualité inclusive des ensembles $\boldsymbol{I}_a$ et $\boldsymbol{I}_b$ pour la règle $a \Rightarrow b$

Pour les données décrites dans le tableau 1.b, l'intensité d'implication vaut 1. Les valeurs entropiques sont $h_1 = 0.918$ et $h_2 = 0.391$. D'où $\tau(a, b) = 0.6035$ et $\Psi(a, b) = 0.77$ alors que $\Pr(b / a) = 0.6666$.

	b	$\neg b$	marge
a	200	400	600
$\neg a$	600	2800	3400
marge	800	3200	4000

(a)

	b	$\neg b$	marge
a	400	200	300
$\neg a$	1000	2400	3400
marge	1400	2600	4000

(b)

	b	$\neg b$	marge
a	40	20	60
$\neg a$	60	280	340
marge	100	300	400

(c)

Tableaux 1. Exemples

Remarquons que la correspondance entre $\varphi(a, b)$ et $\Psi(a, b)$ n'est pas monotone comme le montre l'exemple du tableau 1.c. Dans ce cas, l'intensité d'implication est inférieure à celle du tableau 1.b. Les valeurs entropiques sont $h_1 = 0.918$ et $h_2 = 0.353$. D'où $\Psi(a, b) = 0.78$ alors que $\Pr(b / a) = 0.6666$. Ainsi, alors que $\varphi(a, b)$ a décru de l'exemple précédent à celui-ci, $\tau(a, b)$ et $\Psi(a, b)$ ont crû. Notons cependant, d'une part que la situation inverse est la plus fréquente, et d'autre part que dans les deux cas la probabilité conditionnelle n'a pas changé.

8. EXTENSION AUX VARIABLES ORDINALES, NUMERIQUES ET

D'INTERVALLE

Dans la présentation de l'intensité d'implication, nous nous sommes focalisés sur le traitement des attributs binaires. Cependant, des extensions ont été proposées ces dernières années pour d'autres types de variables. Nous nous contentons de citer ici les références principales auxquelles nous renvoyons pour de plus amples détails.

Une grande partie des travaux a porté sur l'extension du concept d'implication statistique à des variables ordinales : les valeurs $a(i) \in [0,1]$ décrivent des degrés ordonnés d'appartenance ou de satisfaction. Dans ce cas, J. B. Lagrange [16] remplace la valeur observée $q(a, b)$ dans la définition de l'intensité d'implication par *l'indice de propension* défini par :

$$q'(a,\ b) = \frac{\sum_{i \in I} \left(a(i)\ .\ \neg b(i)\right) - \left(n_a n_{\neg b} / n\right)}{\sqrt{\frac{\left(n^2\ s_a^2 + n_a^2\right)\ .\ \left(n^2\ s_{\neg b}^2 + n_{\neg b}^2\right)}{n}}}$$

où $a(i)$ et $\neg\, b(i)$ sont les valeurs prises par l'individu i pour les variables ordinales a et non b ($\neg b(i) = 1 - b(i)$) et s_a et $s_{\neg b}$ les écarts types empiriques de a et non b.

Dans le cas binaire, on retrouve l'intensité d'implication définie paragraphe 2. En effet, si les nombres de modalités de a et b valent 2 alors $n^2\ s_a^2 + n_a^2 = nn_a$, $n^2\ s_{\neg b}^2 + n_{\neg b}^2 = nn_{\neg b}$, et $\sum_{i \in I} a(i)\ .\ \neg b(i) = n_a\ n_{\neg b}\ /\ n$.

Cette solution apportée au cas ordinal est aussi directement applicable au cas des *variables numériques positives,* à condition d'avoir normalisé les valeurs des différentes variables sur I. Lagrange [16] propose une normalisation sur $[0,1]$ en remplaçant les valeurs de la variable a pour chaque individu $i \in \boldsymbol{I}$ par la différence $\mathrm{Max}(a(i)\ ;\ i \in \boldsymbol{I}) - a(i)$.

Pour éviter la normalisation, S. Guillaume [14] a récemment introduit, pour mesurer la tendance implicative entre a et b, l'indice :

$$q''(a,\ b) = \sum_{i \in I} \left(a(i) - \min_{i \in I}\ a(i)\right) \left(\max_{i \in I}\ b(i) - b(i)\right)$$

actualisant ainsi différemment l'indice de propension de J.B.Lagrange.

Nous cherchons actuellement à étendre nos travaux à la recherche de partitions "pertinentes " d'intervalles de valeurs numériques. La situation générique est la suivante : deux variables réelles a et b prennent leurs valeurs au sein de la population sur deux intervalles finis $[a_1, a_2]$ et $[b_1, b_2]$. Par exemple, a représente les poids d'un ensemble de n sujets et b les tailles de ces mêmes sujets. Deux problèmes se posent alors. Peut-on définir des partitions en sous-intervalles adjacents de $[a_1, a_2]$ (resp. $[b_1, b_2]$) afin que la partition la plus fine obtenue respecte au mieux la distribution

statistique des valeurs observées dans $[a_1, a_2]$ (resp. $[b_1, b_2]$) ? Et, peut-on trouver deux partitions l'une de $[a_1, a_2]$ et l'autre *de* $[b_1, b_2]$ constituées de réunions des sous-intervalles adjacents précédents, partitions qui maximisent l'intensité d'implication moyenne des sous-intervalles de l'un sur des sous-intervalles de l'autre et appartenant à ces deux partitions ?

La première question, classique, a conduit à de nombreux travaux dans d'autres cadres que celui exploré ici, en particulier celui de l'apprentissage (voir [25]). Récemment, D. Lahanier-Reuter [17] a proposé un algorithme de partitionnement basé sur l'analyse implicative. Et, nous développons actuellement une méthode basée sur les nuées dynamiques [5] pour traiter conjointement les deux problèmes.

9. CONCLUSION

Nous avons présenté diverses composantes d'une méthodologie de découverte de quasi-implications entre des attributs et des classes d'attributs en insistant, d'une part sur la construction des hiérarchies implicatives et d'autre part sur la mise en œuvre de la mesure de base, l'intensité d'implication, pour des corpus de grande taille. Notons, que pour faciliter les traitements algorithmiques et graphiques toutes les composantes présentées ici ont été intégrées dans un logiciel – C.H.I.C. [4].

Outre les généralisations de l'approche à d'autres attributs que les attributs binaires cités dans le paragraphe ci-dessus, des prolongements sont actuellement en cours dans deux directions.

Dans le cadre de l'ECD mettant en jeu de grands volumes de données, il est courant d'être confronté à des valeurs d'attributs manquantes – ou peu fiables – sur un nombre significatif d'individus. Pour traiter ce problème, une piste consiste à prendre en compte pour l'étude du comportement d'un individu confronté à une valeur manquante sur un attribut les valeurs des autres attributs qui lui sont liés au sens de la quasi-implication.

Un autre problème, plus ouvert, soulevé dans le même cadre est celui de la manipulation d'un nombre élevé d'attributs décrivant les individus. Les algorithmes automatiques classiques de fouille de règles d'association engendrent souvent des ensembles prohibitifs de règles qui nécessitent impérativement pour leur exploitation des post-traitements. La recherche en amont des méthodes d'extraction de classes de "quasi-équivalence " sur l'ensemble des attributs pourrait apporter une simplification à ce problème.

BIBLIOGRAPHIE

[1] AGRAWAL, R., IMIELINSKY, T., SWAMI, A., " Mining association rules between sets of items in large databases *", Proc. of the ACM SIGMOD Conf. on Management of Data*, ACM Press, 1993, p. 207-216.

[2] BAILLEUL, M., GRAS, R., " L'implication statistique entre variables modales ", *Mathématique, Informatique et Sciences humaines* 128, 1995, p. 41-57.

[3] BERNARD, J.-M., POITRENAUD, S., " L'analyse implicative bayesienne d'un questionnaire binaire : quasi-implications et treillis de Galois simplifié ", *Mathématiques, Informatique et Sciences humaines* 147, 1999, p. 25-46.

[4] COUTURIER, R., " Traitement de l'analyse statistique dans CHIC ", *Actes des Journées sur La Fouille dans les Données par la Méthode d'Analyse Statistique Implicative*, IUFM de Caen, 2000, p. 33-50.

[5] DIDAY, E., " Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes ", Thèse d'état, Université de Paris VI, 1972.

[6] FLEURY, L., BRIAND, H., PHILIPPE, J., DJERABA, C., " Rule evaluation for knowledge discovery in databases ", *Proc. of the 6th Conf. on Database and Expert System App.*, Elsevier Sc., 1995, p. 405-414.

[7] GRAS, R., *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse d'état, Université de Rennes I, 1979.

[8] GRAS, R., LARHER, A., " L'implication statistique, une nouvelle méthode d'analyse de données ", *Mathématique, Informatique et Sciences humaines* 120, 1992, p.□-31.

[9] GRAS, R. et coll., *L'implication Statistique*, Grenoble, La Pensée Sauvage, 1996.

[10] GRAS, R., RATSIMBA-RAJOHN, H., " Analyse non symétrique de données par l'implication statistique ", *RAIRO-Recherche Opérationnelle* 30-3, Paris, Elsevier, 1996, p. 217-232.

[11] GRAS, R., BRIAND, H., PETER, P., PHILIPPE, J., " Implicative statistical analysis ", *Proc. of the Congress of the Int. Fed. of Classification Soc.,* Springer-Verlag, 1997, p. 412-419.

[12] GRAS, R., KUNTZ, P., COUTURIER, R., GUILLET, F., " Une version entropique de l'implication statistique pour des corpus de grande taille ", *Extraction des Connaissances et Apprentissage* 1-1/2, Hermès, 2001, p. 69-80.

[13] GUILLAUME, S., GUILLET, F., PHILIPPE, J., " Improving the discovery of association rules with intensity of implication ", *Proc. of the 2nd Eur. Symp. on Principles of Data Mining and Knowledge Discovery, Lecture Notes in Comp. Sc. 1510*, Springer, 1998, p. 318-327.

[14] GUILLAUME, S., *Traitement des données volumineuses - Mesures et algorithmes d'extraction de règles d'association et de règles ordinales*, Thèse de l'Université de Nantes, 2000.

[15] KUNTZ, P., GUILLET, F., LEHN, R., BRIAND, H., " A user-driven process for mining association rules ", *Proc. of Principles of Data Mining and Knowledge Discovery*, *Lecture Notes in Art. Int. 1910*, Springer, 2000, p. 483-489.

[16] LAGRANGE, J.B., " Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire aux réponses modales ordonnées ", *Revue de Statistique Appliquée* XLVI-1, Paris, I.H.P., 1998, p. 71-93.

[17] LAHANIER-REUTER, D., " Étude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire", Thèse de l'Université de Rennes 1, 1998.

[18] LERMAN, I.C., *Les bases de la classification hiérarchique*, Paris, Gauthier-Villars, 1970.

[19] LERMAN, I.C., *Classification et analyse ordinale des données*, Paris, Dunod, 1981.

[20] LERMAN, I.C., GRAS, R., ROSTAM, H., " Élaboration et évaluation d'un indice d'implication pour des données binaires I et II ", *Mathématiques et Sciences humaines* 74, 1981, p. 5-35 et 75, 1981, p. 5-47.

[21] LOEVINGER, J., " A systematic approach to the construction and evaluation of tests of abilities ", *Psychological Monographs* 61(4), 1947.

[22] SHANNON, C.E., WEAVER, W., *The mathematical theory of communication*, Univ. of Illinois Press, 1949.

[23] SUZUKI, E., ZYTKOW, J.-M., " Unified algorithm for undirected discovery of exception rules ", *Proc. of Principles of Data Mining and Knowledge Discovery, Lecture Notes in Art. Int. 1910*, Springer, 2000, p. 169-180.

[24] WALD, A., WOLFOWITZ, J., " Statistical tests based on permutations of the observations ", *Ann. Math. Stat.*, 15, 1944.

[25] ZIGHED, D., RAKOTOMALALA, R., *Graphes d'induction - Apprentissage et data mining*, Hermès, 2000.