EXTENSION DE L'ANALYSE STATISTIQUE IMPLICATIVE AU CAS CONTINU DE L'ESPACE DES SUJETS

Régis GRAS[1], Jean-Claude REGNIER[2]

TITLE

Extension of the Statistical Implicative Analysis for continuous subject spaces.

RÉSUMÉ

L'Analyse Statistique Implicative (A.S.I.) classique permet d'extraire des règles et des méta-règles entre des variables de nature variée à partir de données d'une population discrète et finie. Nous envisageons ici l'extension de cette méthode à une population continue sur laquelle est définie une distribution de probabilité donnée. Nous obtenons des indices de qualité des règles extraites de variables booléennes sur une telle population. Nous illustrons cette nouvelle extension de l'A.S.I. par des exemples. Nous montrons que la restriction au cas classique de cette extension au cas continu est valide.

Mots-clés *: Densité, fonction Gamma, intensité d'implication, indice d'implication, loi de Gauss, mesure de probabilité, règle d'association.*

ABSTRACT

The classical Statistical Implicative Analysis (SIA) allows extracting various rules and meta-rules between variables from a discrete and finite population. We present in this paper an extension of this method to a continuous population on which a probability distribution has been defined. We obtain some measures of quality for the rules extracted from boolean variables coming from such a population. We illustrate this new extension of SIA by several examples. We show that the restriction of this extension to the classical situation is valid.

Keywords *: Density, Gamma function, implication intensity, measure of implication, Gaussian law, probability measure, association rule*

1 Introduction

L'Analyse Statistique Implicative (A.S.I.) est une méthode d'analyse des données issues du croisement d'une population d'objets-sujets E et d'un ensemble de variables quelconques (booléennes, intervalles, floues, vectorielles, …). Cette méthode vise à attribuer une valeur d'intensité ou de qualité à des règles d'association (tendancielle ou implicative) du type : *« si la variable a est observée dans la population alors la variable b a tendance à être également observée »*. Elle fournit pour cela un cadre statistique permettant d'aborder, voire de traiter à travers leur quantification, des

[1] Régis Gras, Ecole Polytechnique de l'Université de Nantes, Équipe Connaissance et Décision, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, Site de la Chantrerie, rue C.Pauc, BP 44306, Nantes cedex 3, e-mail : regisgra@club-internet.fr,

[2] Jean-Claude Régnier, Laboratoire UMR 5191 ICAR – Université Lumière de Lyon –Lyon2, 86 rue Pasteur, 69635 LYON Cedex 07, e-mail : jean-claude.regnier@univ-lyon2.fr

relations causales entre phénomènes. Ce cadre s'appuie sur une mesure de probabilité - l'intensité d'implication - de l'invraisemblance de la liaison entre les variables dans une hypothèse d'indépendance a priori de ces variables[3]. Cette approche est comparable à celle pratiquée et discutée par (Lagrange, 1998), (Lerman,1981), (Lerman et Azé, 2004), (Lerman et Guillaume, 2010) pour calculer la mesure symétrique d'une règle d'association de type ressemblance ou implication. Suite aux travaux précurseurs de (Agrawal et al., 1993) et en raison de l'importance des applications en Extraction de Connaissances dans les Données (ECD), de nombreux algorithmes ont été développés ces dernières années pour extraire automatiquement des règles d'association (par exemple (Hipp et al., 2000), et autres références dans (Guillet et Hamilton, 2007), et (Lenca et al., 2011, 2007).

Rappelons que les objectifs et les travaux de l'A.S.I. ne s'arrêtent pas à la détermination d'un indice d'intensité d'implication et à l'extraction de règles. En effet, celles-ci s'organisent à travers une structuration de l'ensemble des variables à savoir :

- des graphes de règles (selon des chemins)
- des hiérarchies orientées de méta-règles (selon des classes ordonnées)

qui conduisent à des interprétations en termes de concepts, typiques de structures comportementales cachées de sujets de E. De plus, la notion de contribution de sujets ou de descripteurs de sujets à un chemin ou à une classe de règles permet, en retour, de définir une structure topologique sur l'ensemble E, pour y «dessiner une topologie du sens »[4]. Ainsi, et cela paraît essentiel, à l'instar des méthodes factorielles et classificatoires, l'A.S.I. s'intéresse principalement à la recherche et aux représentations d'une structure plutôt que de s'arrêter à l'extraction de séries de règles, le sens n'émergeant que de la structure[5] où le « tout » va au-delà de l'ensemble des « parties ».

Nous avons jusqu'alors considéré et conceptualisé l'analyse implicative dans le cas du croisement de E, discret, fini, de cardinal n, avec un ensemble de variables V de cardinal p. Nous avons abordé, par ailleurs, le cas des variables continues dans le cadre des variables-intervalles (Gras et al., 2001), (Gras et Kuntz, 2008) ; (Gras et Régnier, 2009) ; mais aussi, par exemple, (Muhlenbach et Rakotomala, 2005) et l'équipe de recherche autour de E. Diday. En général, le cas des variables continues est traité dans le cadre de la régression linéaire. Or, il existe des situations réelles où l'ensemble E des objets-sujets est lui-même un ensemble continu, en tant que sous-ensemble de R, constitué de « formes » quantifiables (par exemple, des couleurs, des sons, des odeurs, des sensations,…), plutôt que de sujets indexés de 1 à n. On rencontre également, par exemple en physique des particules, de situations où les objets appartiennent à des

[3] Nous pourrions rejoindre en cela René Thom (Thom R., 1980) qui écrit : « …le problème n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens *pour nous, ce qui est surprenant dans l'ensemble des faits*. Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer ».

[4] P. Gaudin, « Y a-t-il de la non-linéarité dans la sémantique ? » dans *Émergence, complexité et dialectique*, Odile Jacob, Paris, p. 279-288)

[5] « …le tout ne se compose de rien d'autre que de ses parties, et pourtant il présente en tant que tout des propriétés n'appartenant à aucune de ses parties. Autrement dit, dans le passage non additif, non linéaire des parties au tout, il y a apparition de propriétés qui ne sont d'aucune manière précontenues dans les parties et ne peuvent donc s'expliquer par elles » Sève L. et al, *Emergence, complexité et dialectique,* Odile Jacob, Paris, p.58, 2005.

espaces infinis dénombrables assimilables à des espaces continus (par ex. mouvements browniens de molécules, marche aléatoire sur une droite,…).

Dans ce texte, nous élargissons notre approche traditionnelle en introduisant cette classe nouvelle de situations dans laquelle l'espace E est continu. Nous montrerons que la restriction au cas fini discret, de cette modélisation du cas continu coïncide, avec l'approche classique de l'ASI et des mesures associées. Dans la littérature – les réseaux bayesiens, par exemple – la prise en compte de telles situations (espace des sujets continu) est, à notre connaissance, escamotée par une discrétisation de cet espace. C'est également le cas dans les autres recherches portant sur les règles d'association, qu'elles soient symétriques ou dissymétriques où l'espace des sujets est fini et discret. Or, préserver la structure continue de l'espace de sujets permet de respecter les nuances que la discrétisation risque d'effacer en une hétérogénéité artificielle. Dans *La cohérence du réel* (Gauthier-Villars, 1989), Erwin Lazlo (1989, p.186) affirme que «*Il n'y a pas d'immaculée perception : nous ne voyons la réalité qu'au travers des lunettes de la théorie* ». La réduction continu->discret, à laquelle nous nous sommes nous-mêmes bien souvent livrés, comporte ce risque que, dans la mesure du possible, nous tenterons d'exclure ou de contrôler.

2 - Espace E des sujets continu, muni d'une mesure

Considérons alors un ensemble E, de la droite réelle, une tribu de boréliens[6] $\mathcal{B}$ sur E, munie d'une mesure de probabilité μ et la mesure produit $\nu=\mu x\mu$ sur ExE (par exemple, une aire normalisée), fonction mesurable de densité f à valeurs réelles positives[7].

Considérons deux boréliens A et B de $(E, \mathcal{B}, \mu)$ tels que $\mu(A) \leq \mu(B)$[8], parties où s'observent les réalisations respectives des variables binaires a et b. Associons à A et B respectivement, deux boréliens aléatoires, indépendants a priori, X et Y définies sur $(E, \mathcal{B}, \mu)$ telles que les rectangles aléatoires $Xx\,E$ et $Ex\overline{Y}$ aient des ν–mesures égales aux ν–mesures respectives de AxE et $Ex\overline{B}$. On cherche, comme dans la forme classique de l'ASI, à estimer la force de la liaison entre a et b permettant d'assurer avec confiance que l'observation de a est généralement accompagnée de celle de b, c'est-à-dire que la règle $a \Rightarrow b$ est de bonne qualité. Autrement dit, il nous faut estimer la petitesse attendue de la mesure $\nu[(XxE) \cap (Ex\overline{Y})]$ de l'ensemble où l'on rencontre les situations contre-exemples de l'implication et la comparer à la valeur observée $\nu[(AxE) \cap (Ex\overline{B})]$. Dans le cas classique, E était discret et la mesure définie sur E était une mesure cardinale. De ce fait, nous devions comparer le cardinal de l'ensemble des contre-exemples $A \cap \overline{B}$ à celui de l'ensemble aléatoire $X \cap \overline{Y}$, dans l'hypothèse où X et Y auraient été des parties indépendantes.

L'introduction d'une densité f permet d'ouvrir plus largement les applications car il est alors possible de pondérer de façon différente les instances $x \in E$ ou les sous-ensembles d'instances des variables. Par exemple, on pourra pondérer continûment selon un descripteur-objet de la contingence. Nous verrons d'ailleurs plus loin que la solution proposée dans le cadre des variables continues où nous avons discrétisé la

[6] Un borélien sur R est généralement un intervalle ou une réunion d'intervalles de R.

[7] On peut aborder la situation de façon différente, mais moins intuitive, en écrivant qu'étant donné la mesure λ de Borel-Stieljes sur $\mathbb{R}^2$ et la fonction mesurable f, l'application : $\forall F \in \mathcal{B}^2$, $F \rightarrow \int_F f\,d\lambda$ est une mesure positive μ telle que $\lambda(F) = 0 \Rightarrow \mu(F) = 0$.

[8] Nous notons dans un souci de simplification des notations : $\mu(A)$ au lieu de $\nu\{(x, y) \in Ax\,E\}$

distribution, est transposable dans celui d'un tel ensemble E munie d'une mesure μ (Gras et al., 2001).

La mesure de la variable aléatoire, $\nu[(X \times E) \cap (E \times \overline{Y})]$, ensemble des contre-exemples à la règle a⇒b, est une variable aléatoire de $(E^2, \mathcal{B}^2, \nu)$, à valeurs réelles positives, définie par l'application h de ExE dans [0,1] et telle que $\forall x \in [0,1]$, $\{\omega : h(\omega)=x\} \in \mathcal{B}$. Sa réalisation empirique est $\mu[(A \times E) \cap (E \times \overline{B})]$.

Appliquant le théorème de Fubini, on obtient la mesure :

$\nu[(X \times E) \cap (E \times \overline{Y})] = [\nu\{(x,y)/x \in X, y \in \overline{Y}] = [\int_X \left[\int_{\overline{Y}} f(y/x)dy\right] f_1(x)\, dx]$ (1)

expression dans laquelle :

- f est la densité de la mesure produit μ sur ExE ;
- f_1 (resp. f_2) est la densité marginale sur le premier (resp. 2ème) espace facteur vérifiant donc $f_1(x) = \int_E f(x,y)dy$ (resp. $f_2 = \int_E f(x,y)\, dx$)
- f(./x) est la densité conditionnelle $\frac{f}{f_1}$ sur Y, sachant x

Puisque les parties X et Y sont indépendantes par hypothèse, elles engendrent des sous-tribus elles-mêmes indépendantes et les mesures-images sur E ont des densités conditionnelles vérifiant : $f = f_1.f_2$. Ce qui implique que f(./x) et f(./y) restreintes aux boréliens engendrés par X et Y, coïncident respectivement avec les densités $f_1$et f_2 sur E et que, dans ce cas, (1) s'écrit :

$[\nu[(X \times E) \cap (E \times \overline{Y})] = [\int_X \left[\int_{\overline{Y}} f(y/x)dy\right] f_1(x)dx] = [\int_{\overline{Y}} f_2(y)dy \int_X f_1(x)dx] =$
$\nu[(X \times E)]. \nu[(E \times \overline{Y})]$

La réalisation empirique ou valeur contingente de la mesure aléatoire $[\nu[(X \times E) \cap (E \times \overline{Y})]$ est : $\nu[(A \times E) \cap (E \times \overline{B})] = \int_A \left[\int_{\overline{B}} f(y/x)dy\right] f_1(x)dx$ (2)

Pour alléger la notation, dans le cas général où E=R et se rapprocher des formes utilisées dans le cas classique, nous utiliserons les notations abusives suivantes : $\mu(A \cap \overline{B})$ au lieu de $\nu[(A \times E) \cap (E \times \overline{B})]$ et $\mu(X \cap \overline{Y})$ au lieu de $[\nu[(X \times E) \cap (E \times \overline{Y})]$. Cet abus d'écriture n'en est plus un si la mesure μ est normalisée (i.e. $\mu(E) = 1$) et si l'on considère la mesure projetée sur chacune des dimensions E.

On a donc, par construction des parties aléatoires X et Y, l'égalité des espérances de leurs mesures aléatoires :

$\mathbb{E}[\mu(X)]$ [9]$= \mu(A)$; $\mathbb{E}[\mu(\overline{Y})] = \mu(\overline{B})$; $\mathbb{E}[\mu(X \cap \overline{Y})] = \mu(A \cap \overline{B})$

La différence aléatoire $\mu(A \cap \overline{B}) - \mu(X \cap \overline{Y})$, sous l'hypothèse d'indépendance a priori de X et Y, est un indicateur de qualité de la règle a ⇒ b. Plus la différence est forte entre la valeur contingente $\mu(A \cap \overline{B})$ et la valeur attendue $\mu(X \cap \overline{Y})$, sous cette hypothèse plus forte sera la présomption de relation implicative, voire causale, de a sur b. Plus précisément, dès lors que la loi de $\mu(X \cap \overline{Y})$ sera connue, c'est la valeur de :

$1 - P[\mu(X \cap \overline{Y}) \leq \mu(A \cap \overline{B})]$ [10] qui sera retenue comme premier indicateur de la qualité de la règle. Plus la probabilité de l'inégalité est faible, meilleure est la qualité de la règle. Notons que $P[\mu(X \cap \overline{Y}) \leq \mu(A \cap \overline{B})]$ est la valeur de la fonction de répartition de $\mu(X \cap \overline{Y})$ au point $\mu(A \cap \overline{B})$.

Cependant, cette valeur est toute relative et ne fournit pas de référence par rapport à une échelle de probabilité, comme nous l'obtenons dans les différents cas que nous avons traités avec divers types de variables et un ensemble discret fini de la population des sujets. Toutefois, lorsque l'on connaît la densité f de la mesure μ, nous verrons à travers deux exemples qu'il est possible de déterminer cette échelle de probabilité. Nous reprenons donc la démarche que nous avons adoptée dans le cas discret.

---

[9] $\mathbb{E}[\mu(X)]$ (resp. $\mathbb{E}[\mu(\overline{Y})]$, resp. $\mathbb{E}[\mu(X \cap \overline{Y})]$) est l'espérance mathématique de la variable $\mu(X)$ (resp. $\mu(\overline{Y})$, resp. $\mu(X \cap \overline{Y})$ de densité f_1 (resp. f_2, resp. f)

[10] P est un probabilité définie par $P(F) = \nu(\{\omega : h(\omega) \in F\}$ où F est un événement de ExE

Si la loi de $\mu(X\cap \overline{Y})$ le permet (par exemple par une approximation gaussienne), on s'intéressera plutôt à la variable aléatoire $\mu(X\cap \overline{Y})$ centrée réduite

$$Q(X \cap \overline{Y}) = \frac{\mu(X\cap\overline{Y}) - E[\mu(X\cap\overline{Y})]}{\sqrt{[Var(\mu(X\cap\overline{Y})]}} \quad (3)$$

dont la loi de probabilité sera approximativement celle de $\mathcal{N}(0,1)$ et dont l'observation $q(a,\overline{b})$ fournira un indice d'implication de base, dans ce cas continu. Il mesurera l'écart normalisé entre la contingence et l'indépendance

$$q(a,\overline{b}) = \frac{\mu(A\cap\overline{B}) - E[\mu(X\cap\overline{Y})]}{\sqrt{[Var(\mu(X\cap\overline{Y})]}}$$

Or par hypothèse :
$\mu(X \cap \overline{Y}) = \mu(X).\mu(\overline{Y})$ et $[\mathcal{E}(X \cap \overline{Y}) = \mathcal{E}(\mu(X)).\mathcal{E}(\mu(\overline{Y})) = \mu(A).\mu(\overline{B})$
L'indice d'implication prend alors la forme :

$$q(a,\overline{b}) = \frac{\mu(A\cap\overline{B}) - \mu(A).\mu(\overline{B})}{\sqrt{[(Var(\mu(X\cap\overline{Y})]}} \quad (4)$$

expression dans laquelle $\mu(X)$ et $\mu(\overline{Y})$ étant indépendantes de même que leurs carrés puisque l'application $U\rightarrow U^2$ est mesurable [Métivier, 1968]:
$Var(\mu(X \cap \overline{Y}) = Var(\mu(X).\mu(\overline{Y}) = \mathcal{E}[\mu(X\cap\overline{Y})^2] - \mathcal{E}[\mu(X\cap\overline{Y})]^2 =$
$\mathcal{E}[\mu(X)^2.\mu(\overline{Y})^2] - \mathcal{E}[\mu(X).\mu(\overline{Y})]^2 = \mathcal{E}[\mu(X)^2.\mu(\overline{Y})^2] - \mathcal{E}[\mu(X)]^2.\mathcal{E}[\mu(\overline{Y})]^2$
$\mathcal{E}[\mu(X)^2.\mu(\overline{Y})^2] - \mu(A)^2.\mu(\overline{B})^2$
L'intensité d'implication, dans le cas d'une approximation gaussienne devient alors :

$$\varphi(a,b) = 1 - P[Q(X \cap \overline{Y}) \leq q(a,\overline{b})] \text{ soit encore } \varphi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\overline{b})}^{+\infty} e^{-\frac{t^2}{2}}\, dt \quad (5)$$

Remarque 1

On notera que la détermination de l'indice d'implication $q(a,\overline{b})$ et, par suite, de l'intensité d'implication, ne nécessite que la connaissance de la densité f de μ.

Remarque 2

Si la densité n'est pas connue, la méthode des moments permet, en pratique, de déterminer empiriquement les moments théoriques dont nous avons besoin dans le calcul de l'indice d'implication, à savoir la moyenne et la variance.

Remarque 3

Dans le cas où la densité est uniforme sur ExE, il est aisé de définir la loi de la variable $\mu(X\cap \overline{Y})$ puisqu'il y a proportionnalité entre « aire » et probabilité. On le montrera dans l'exemple 3. Dans le cas général, une étude plus fine à partir de la densité f conduit à déterminer quel « étonnement statistique » est associé à l'observation de la différence δ entre $\mu(A\cap\overline{B})$ et $\mu(X\cap \overline{Y})$ et donc quelle qualité recèle la règle implicative entre les variables a et b.

Remarque 4

Le traitement du tableau de données ExV peut donc, en théorie, se poursuivre et permettre la construction du graphe implicatif représentant la structure implicative des règles entre les variables. La hiérarchie cohésitive représentera la structure des méta-

règles extraites de ExV. De plus, il sera, comme dans la théorie générale des populations discrètes (ou discrétisées) examiner les contributions des sous-ensembles réels sur lesquels se réalisent les variables.

3 Trois exemples numériques

3.1 Exemple 1

Dans un centre de formation à l'œnologie, on a établi un étalonnage des compétences d'une grande population de buveurs de vins. Il s'exprime par une valeur numérique qui va de l'incompétence totale à la grande compétence modélisable par une variable gaussienne centrée réduite. E est alors « population » supposée infinie et continue d'une échelle de compétences (injection de l'ensemble des compétences sur une population de buveurs)[11].

Sur un stand d'une foire aux vins, un producteur propose, au cours de plusieurs journées de dégustation, à ses dégustateurs, un questionnaire relatif aux critères de sa production de St Emilion. Plusieurs critères sont évaluables relativement aux arômes et aux qualités de vieillissement de son vin. On ne retient, dans cet exemple, que deux critères : la complexité du vin et sa capacité à bien vieillir. Ces deux variables sont binaires : le vin est jugé complexe (a=1) ou non(a=0), le vieillissement est attendu (b=1) ou pas(b=0). La population humaine interrogée présente une certaine compétence a priori que l'on peut supposer relever d'une loi gaussienne $\mathcal{N}(0,1)$.

Soit A (resp.B) l'ensemble des intervalles réels de compétences associées à la population sur laquelle a (resp.b) est vérifiée. $A \cap \overline{B}$ est la population qui émet une opinion positive sur la complexité et négative sur le vieillissement.

L'étude de l'implication de la variable a sur la variable b devrait se faire par calculs d'intégrales gaussiennes de μ(A) (resp. μ(B)) sur chacun des intervalles de compétences et d'en faire la somme. Afin de simplifier ces calculs, nous supposerons - ce qui n'enlève d'aucune façon le caractère de généralité - que A et B sont réduits à un seul intervalle de E.

$$\boldsymbol{\mu(A)} = \frac{1}{\sqrt{2\pi}}\int_A e^{-\frac{x^2}{2}}\,dx \;;\; \boldsymbol{\mu(\overline{B})} = \frac{1}{\sqrt{2\pi}}\int_{\overline{B}} e^{-\frac{y^2}{2}}\,dy \;;\; \mu(A \cap \overline{B}) = \frac{1}{\sqrt{2\pi}}\int_{A\cap\overline{B}} e^{-\frac{u^2}{2}}\,du$$

Nous disposons d'une mesure de la qualité de la relation implicative entre les variables a et b par l'indicateur 1 - Pr [μ(X∩ $\overline{Y}$) ≤ μ(A∩$\overline{B}$)].

En effet, Pr [μ(X∩ $\overline{Y}$) ≤ μ(A∩$\overline{B}$)] = $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mu[A\cap\overline{B}])} e^{-\frac{u^2}{2}}\,du$

[11] Une situation comparable est rencontrée dans la théorie des tests où un paramètre d'intérêt $\boldsymbol{\theta}$ dénote la capacité ou le niveau d'aptitude de sujets. Il est coutumier de considérer que cette capacité soit distribuée normalement (par ex. $\mathcal{N}(0,1)$) dans la population de répondants à des items d'un test. Ce paramètre θ, lorsqu'il est connu par estimation permet de calculer la probabilité $P_i(\theta)$ que le sujet de capacité θ réussisse à l'item i du test. La théorie des réponses aux items (I.R.T.) est fondée sur l'estimation de cette capacité sur la base des réponses aux variables items.

Revenant au calcul proposé par la formule (4), celui de $Var(\mu(X \cap \overline{Y})$ est plus complexe. Formellement, cette variance se développe en :

$$\mathcal{E}[\mu(X \cap \overline{Y})^2] - \mathcal{E}[\mu(X \cap \overline{Y})^2] = \mathcal{E}[\mu(X)^2].\ \mathcal{E}[\mu(\overline{Y})^2] - \mu(A)^2.\mu(\overline{B})^2$$

Pour calculer, $\mathcal{E}[\mu(X)^2]$, il nous faut connaître la densité de $\mu(X)^2$ connaissant celle de $\mu(X)$ qui est f_1. Puisque f_1 est la densité de la loi normale $\mathcal{N}(0,1)$, nous pouvons effectuer le calcul de l'intégrale I représentant l'espérance du carré de la mesure de X:

$\mathcal{E}[\mu(X)^2] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}}\,dt = \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} t^2 e^{-\frac{t^2}{2}}\,dt = \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} v^{\frac{1}{2}} e^{-\frac{v}{2}}\,dv$ en posant $v=t^2$

Puis $I = \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} 2\sqrt{2}u^{\frac{1}{2}}e^{-u}\,du = \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} 2\sqrt{2}u^{\frac{1}{2}}e^{-u}\,du = \frac{2}{\sqrt{\pi}}\int_{0}^{+\infty} u^{\frac{1}{2}}e^{-u}\,du$

Rappelons la propriété de la fonction Gamma:

$\Gamma(x) = \int_{0}^{+\infty} v^{x-1}e^{-v}\,dv$ et $\Gamma(x+1) = x\,\Gamma(x)$,

De même $\mathcal{E}[\mu(\overline{Y})^2] = 1$ et, par suite : $Var(\mu(X \cap \overline{Y}) = 1 - \mu(A)^2.\mu(\overline{B})^2$

La relation (4) devient alors : $q(a,\bar{b}) = \frac{\mu(A\cap\overline{B}) - \mu(A).\mu(\overline{B})}{\sqrt{1-\mu(A)^2.\mu(\overline{B})^2}}$ (6)

Pour illustrer numériquement ce calcul, choisissons : A = [-1 ; 2,6], B = [-2 ; 2,5] deux intervalles fermés de E Nous constatons que $A\cap\overline{B}$ =]2,5 ; 2,6]. Les mesures de ces sous-ensembles sont, en se référant à la loi $\mathcal{N}(0,1)$:

$$\mu(A) = \frac{1}{\sqrt{2\pi}}\int_{-1}^{2,6} e^{-\frac{x^2}{2}}\,dx = 0,836$$

$$\mu(B) = \frac{1}{\sqrt{2\pi}}\int_{-2}^{2,5} e^{-\frac{x^2}{2}}\,dx = 0,9710402 \approx 0,97$$

$$\mu(A \cap \overline{B}) = \frac{1}{\sqrt{2\pi}}\int_{2.5}^{2,6} e^{-\frac{x^2}{2}}\,dx = 0,001$$

Notons qu'alors $Pr[\mu(X \cap \overline{Y}) \le \mu(A \cap \overline{B})] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mu(A\cap\overline{B})} e^{-\frac{x^2}{2}}\,dx = 0,501$ .Le premier indicateur de qualité est : $1 - Pr[\mu(X \cap \overline{Y}) \le \mu(A \cap \overline{B})] = 0,499$

Nous pouvons calculer l'intensité d'implication correspondante à l'aide de la formule (6) et de $q(a,\bar{b})$:

$$q(a,\bar{b}) = \frac{\mu(A \cap \overline{B}) - \mu(A)\mu(\overline{B})}{\sqrt{1-\mu(A)^2\,\mu(\overline{B})^2}} = \frac{0,00154848 - (0,83668356)(0,9710402)}{\sqrt{1-(0,83668356)^2\,(0,9710402)^2}} = -0,02268837$$

$$\varphi(a;b) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{q(a,\bar{b})} \exp(-\frac{t^2}{2})dt = 1 - 0,49094943 = 0,50905057$$

Les calculs ci-dessus indiquent que la complexité est un jugement qui conduit plutôt vers la capacité à vieillir mais cette relation est faible.

3.2 Exemple 2

Dans un souci didactique, afin d'illustrer le fonctionnement d'une telle estimation, nous simulons un tirage au hasard des parties aléatoires X et Y, puis en tirons le calcul d'une autre estimation de l'intensité d'implication. Nous proposons un calcul « à la main » qui exhibe tous les pas de l'expérience aléatoire qui définit notre traitement du cas continu.

On fait varier aléatoirement sur E les segments X et Y de mêmes μ-mesures en moyenne respectivement que A et B, soit 0,8400 et 0,9710. Par exemple X = [-1,00289; 2,87112] et Y = [2,12038; 2,22581] auquel correspond $X \cap \overline{Y}$ = [-1,00289; 2,12038[dont la μ–mesure est 0,01096813 . On calcule, à chaque étape,

$\mu(X \cap \overline{Y})$ par : $\frac{1}{\sqrt{2\pi}} \int_{X \cap \overline{Y}} e^{-\frac{u^2}{2}} du$, puis $[\frac{1}{\sqrt{2\pi}} \int_{X \cap \overline{Y}} e^{-\frac{u^2}{2}} du]^2$ et enfin sa moyenne sur un certain nombre de répétitions. On obtient une estimation de $q(a, \bar{b})$ par la formule (4) et enfin celle de l'intensité d'implication par la formule (5).

A la main, en vue d'une estimation de l'intensité d'implication, nous avons procédé à 10 calculs comme l'indique le tableau ci-dessous, en choisissant au hasard les intervalles aléatoires respectant la contrainte $\mu(X) \leq \mu(Y)$.

TABLEAU 1 – *échantillons d'intervalles*

Tirages	X μ(X)=0.84	Y μ(Y)=0.97	X et nonY	μ(X et nonY)
1	[-1,06884 ; 2,10999]	[-1,88286; 3,63162]	vide	0
2	[-1,00289; 2,87112]	[2,12038; 2,22581]	[-1,00289; 2,12038[	0,01096813
3	[-2,31536; 1,03770]	[-2,06711; 2,30303]	[-2,31536; -2,06711 [	0,00906513
4	[-3,12126 ; 0,99817]	[-1,88962 ; 3,24110]	[-3,12126 ; -1,88962[	0,02850424
5	[-1,98820 ; 1,09570]	[-1,88083 ; 4,56158]	[-1,98820 ; -1,88083[	0,006602909
6	[-2,64815 ; 1, 01123]	[-1, 91166; 2,87192]	[-2,64815 ; -1,91166[	0,023913423
7	[-1,60406; 1,25]	[-1,880791; +∞[	vide	0
8	[-2,06108 ; 1,07874]	[-1,92155 ; 2,78590]	[-2,06108 ; -1,92155[	0,007683136
9	[-2,76884; 1,0060]	[-1,99719; 2,45226]	[-2,76884; -0,99719 [	0,02008914
10	[-1,95161 ; 1,10533]	[-1,92215 ; 2,78136]	[-1,95161 ; -1,92215]	0,0018012

On en tire : MOY[$\mu(X \cap \overline{Y})$]= 0, 01086273 ; [MOY($\mu(X \cap \overline{Y})$)]^2= 0, 000209627

Puis par la formule (6): Est $(q(a, \bar{b})) = -0{,}02268837$

d'où l'estimation de l'intensité d'implication : Est (φ(a,b)) = 0,50061775

Remarquons que cette valeur est très voisine de celle obtenue par un calcul direct selon la formule (5), ce qui conforte l'approche simulée en cas de calculs complexes de la variance lorsque la loi de probabilité μ est inconnue.

3.3 Exemple 3

L'espace E est un segment fermé de la demi-droite réelle positive de longueur 10^6. Ce segment est muni de la restriction à E de la mesure de Borel-Stieltjes qui prolonge la fonction longueur. La densité de probabilité associée g, pour tout t réel t→t est uniforme, égale à $\frac{1}{10^6}$ et telle que μ(E)= 1. Les sous-ensembles A, B, X et Y sont des segments fermés :

A = [0 ; 400000] ; B = [100000 ; 600000] ; $A \cap \overline{B}$ = [0 ; 100000[

$\mu(A)= \int_0^{400000} g(x)dx= 0{,}40$; de même $\mu(B)= \int_{100000}^{600000} g(x)dx= 0{,}50$;

$\mu(A \cap \overline{B})= \int_0^{100000} g(x)dx = 0{,}10.$

Par un choix au hasard respectant l'uniformité de g retenons X = [150000 ; 550000] et Y = [300000 ; 800000], on obtient :

$X \cap \overline{Y}$= [150000 ; 300000[et $\mu(X \cap \overline{Y}) = 0{,}15$.

En simulant un tel tirage 1000 fois, on obtiendrait une valeur moyenne estimation de $\mathbb{E}[\mu(X \cap \overline{Y})^2]$ puis celle de $Var(\mu(X \cap \overline{Y})$ et enfin celle de $q(a, \overline{b})$ et de φ(a,b).

En utilisant la démarche précédente de l'exemple 1 et la formule (6), on obtient successivement : $\mathbb{E}[\mu(X)^2] = \int_0^{10^6} t^2.\frac{1}{10^6}dt = \frac{1}{3} = \mathbb{E}[\mu(\overline{Y})^2]$

puis $\mathbb{E}[\mu(X \cap \overline{Y})^2] - \mathbb{E}[\mu(X \cap \overline{Y})]^2 = \mathbb{E}[\mu(X)^2].\ \mathbb{E}[\mu(\overline{Y})^2] - \mu(A)^2.\mu(\overline{B})^2 = \frac{1}{9} - 0{,}04 =$ 0,071, d'où $q(a, \overline{b}) = - 0{,}376$ et φ(a,b) = 0,644

4 Espace continu et discrétisable

Nous nous inspirons ici de ce que nous avons fait pour discrétiser les variables continues en variables intervalles fermés ou semi-ouverts (Gras et al., 2001). On considère ici un espace des sujets-objets constitué de n intervalles $I_1, I_2, \ldots, I_n$ finis, munis chacun d'une densité de mesure. Soit $g_1, g_2, \ldots g_n$ les densités respectives. Une telle situation peut être rencontrée si l'espace E du § 2 est discrétisé en n intervalles fermés ou semi-ouverts sur lesquels serait définie la restriction à ces intervalles de la densité f. A titre d'exemple, citons le cas où les objets constituant E seraient suffisamment homogènes pour que l'on considère que la relation de chacun d'entre eux est représentative d'une classe relativement à des variables attributs.

Soit deux variables a et b et l'étude de l'implication a⟹b. Soit A et B les supports respectifs de a et b dans E. Ces supports ont pour traces selon les n intervalles les sous-ensembles respectifs $A_1, A_2, \ldots, A_n$ et $B_1, B_2, \ldots, B_n$. Ces sous-ensembles admettent les mesures respectives $\mu(A_i \cap \overline{B_i})$ pour i= 1, 2, …n, calculées comme dans le cas général du § 2 à partir des rectangles projetés sur E et mesurés par les densités respectives données :

$$\mu(A \cap \overline{B}) = \sum_{i=1}^{i=n} \mu(A_i \cap \overline{B_i})$$

On opère de la même façon pour évaluer les mesures des traces des sous-ensembles aléatoires X et Y de mêmes mesures respectives que A et B et pour calculer Espérance et Variance des éléments correspondants.

Notons que cette méthode permet de préserver ou restituer les nuances apportées aux valeurs que prennent les variables, numériques ou non, grâce aux mesures différentes des sous-ensembles qui pondèrent ces valeurs. En ce sens, la méthode est donc plus riche que celle qui consiste, traditionnellement, à découper l'ensemble E démuni d'une mesure.

5 Quelques pistes d'applications

Nous proposons, sans mise en forme complète, la situation concrète relative aux estimations des mesures des bandes passantes perçues par l'oreille suivant que l'on est français (variable "a"), anglais ("b"), espagnol ("c"), allemand ("d") ou russe ("e") ou…. L'espace E des sujets-objets est donc un ensemble continu de fréquences exprimées en hertz (voir les travaux de A. Tomasis sur le sujet, en particulier les courbes que l'on trouve dans son ouvrage (Tomasis, 1977) et sur Google (Wikipedia article relatif à Alfred Tomasis).

A ces bandes passantes correspondent des intervalles continus où l'écoute est optimisée. Les variables, a, b, c, d, e... sont binaires si l'on décide d'affecter la valeur 1 à la variable sur l'intervalle ou les intervalles favorables et 0 sinon. Surviennent alors des questions à traiter dans le cadre A.S.I.: peut-on dire que la qualité de l'oreille française se retrouve dans l'oreille anglaise ou russe ou...? Autrement dit, est-ce que a entraîne b ou c ou... (ou a une propension à...) ? Ceci peut –il expliquer les difficultés de reconnaissance ou d'apprentissage de langue ? Une loi uniforme sur E conviendrait dans un premier temps. Une normalisation gaussienne du grave à l'aigu serait aussi envisageable. A.Tomasis montre, par ses expériences de laboratoire, que l'oreille française « entend essentiellement entre 1000 et 2000 hertz et que l'oreille anglaise inscrit sa sélectivité entre 2000 et 12000 hertz et l'oreille italienne entre 2000 et 4000 hertz ». Ces écarts pourraient expliquer les difficultés des français à percevoir la langue anglaise. Alors que les slaves, disposeraient d'une « facilité à intégrer les langues étrangères » du fait de leur « grande perméabilité auditive ».

Une dernière application dans le domaine de la pédagogie vise à comparer les performances en termes de notes d'une classe d'élèves d'un trimestre au suivant ou d'une année à l'autre. L'échelle de notes constitue un ensemble continu et les variables considérées sont les trimestres ou les années d'observation.

6 Retour au cas classique de l'ASI où E est discret et fini

Choisissons la densité f de telle façon que la répartition soit uniforme dans E, de cardinal fini n, c'est-à-dire telle que f(x) = 1/n pour toute observation x. Nous sommes alors dans le cas de l'ASI classique où les instances (les sujets par exemple) ont le même poids dans la contingence.

Alors : $\mu[(A \cap \overline{B})/A] = \int_{\overline{B}} f(y/x)dy$ soit encore $\frac{n_{a \text{ et } \overline{b}}}{n_a}$

et finalement $\mu(A \cap \overline{B}) = \int_A \frac{n_{a\ et\ \overline{b}}}{n_a} f_1(x)dx = \frac{n_{a\ et\ \overline{b}}}{n_a} . \frac{n_a}{n} = \frac{n_{a\ et\ \overline{b}}}{n}$

De la même façon, on a : $\mu(A) = \frac{n_a}{n}$; $\mu(\overline{B}) = \frac{n_{\overline{b}}}{n}$

L'intensité d'implication devient alors :

$$\varphi(a, b) = 1 - Pr\left[\mu(X \cap \overline{Y}) < \frac{n_{a\ et\ \overline{b}}}{n}\right]$$

Si l'on connaît la loi de tirage aléatoire de X et $\overline{Y}$, on connaît alors la loi de $\mu(X \cap \overline{Y})$ et, par voie de conséquence $\mathbb{E}[\mu(X \cap \overline{Y})]$ et $Var[\mu(X \cap \overline{Y})]$.

Dans le cas classique de l'ASI, ce cardinal suit une loi de Poisson ou une loi binomiale. L'indice d'implication (formule (4)) devient :

$$q(a, \overline{b}) = \frac{\frac{n_{a\ et\ \overline{b}}}{n} - E[\mu(X \cap \overline{Y})]]}{\sqrt{[Var(\mu(X \cap \overline{Y})]}} = \frac{\frac{n_{a\ et\ \overline{b}}}{n} - \mu(A).\mu(\overline{B})}{\sqrt{[Var(\mu(X \cap \overline{Y})]}}$$

Si nous nous plaçons dans le cas où $card(X \cap \overline{Y})$ suit la loi de Poisson de paramètre $\lambda = n.\frac{n_a}{n}.\frac{n_{\overline{b}}}{n}$, on obtient les valeurs suivantes :

$$\mathbb{E}[\mu(X \cap \overline{Y})] = \mathbb{E}\left[\frac{card(X \cap \overline{Y})}{n}\right] = \frac{\lambda}{n} = \mu(A).\mu(\overline{B}) = \frac{n_a n_{\overline{b}}}{n^2}$$

$$Var[\mu(X \cap \overline{Y})] = \frac{1}{n^2} Var\,[card(X \cap \overline{Y})] = \frac{1}{n^2}\lambda = \frac{n_a\, n_{\overline{b}}}{n^3}$$

On obtient alors la formule classique donnant :

$$q(a,\overline{b}) = \frac{n_{a \wedge \overline{b}} - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}}$$

Ainsi, la restriction du cas où nous avons affaire à des espaces continus coïncide avec l'étude déjà faite du cas où l'espace est discret et fini.

Remarque 5

Le cas uniforme et fini envisagé ci-dessus n'est qu'un cas particulier d'une distribution finie où cette fois la distribution n'est pas uniforme. Chaque « sujet » peut contribuer à l'étude avec des pondérations différentes. Il suffit de pondérer les paramètres n_a, n_b et $\boldsymbol{n_{a \wedge \overline{b}}}$ des sommes des pondérations affectant chacun des « sujets » en jeu. C'est, par exemple, le cas où la population analysée est composée de descripteurs regroupant des sujets présentant le même attribut.

7 Conclusion

En nous inspirant de la méthode s'appuyant sur le calcul des probabilités et des statistiques qui nous a permis de définir un indice de qualité de règles implicatives entre variables observées sur une population discrète, nous avons dans cet article défini un tel indice mais cette fois sur une population de sujets-objets continue. Pour cela, nous avons introduit une mesure de probabilité sur l'ensemble continu des contre-exemples aux règles attendues entre des variables binaires et cela en tirant au hasard des supports indépendants. Cette mesure nous permet de mesurer l'écart entre la contingence et l'ensemble des contre-exemples attendus dans cette hypothèse d'indépendance. Nos

illustrations montrent le processus sous-jacent à l'atteinte de la mesure de qualité des règles. Il est bien évident que la démarche du § 2 est assez générale pour être appliquée à d'autres méthodes d'analyse de données et en particulier à l'analyse des similarités au sens de I.C. Lerman, c'est-à-dire définie à partir de *l'algorithme de la vraisemblance du lien (A.V.L.)*

Nous montrons que la restriction de cette mesure au cas discret nous permet de retrouver la mesure de qualité jusqu'alors définie. Il serait maintenant intéressant de mettre en pratique ce nouvel outil sur des exemples tirés de situations réelles, comme celles évoquées dans le paragraphe §5, où la nécessité de conserver le caractère continu pour en maintenir les nuances conduit à éviter la discrétisation qui, elle, risquerait de les faire disparaître.

Nous pensons également que la modélisation probabiliste que nous avons faite dans le §2 pourrait être transposée dans le cas où cette fois ce sont les variables qui sont continues. La donnée d'une loi de probabilité sur chacune ou sur une partie des variables, voire une estimation de leur loi, permettrait de définir un indice de qualité de mesure d'implication du type « si a alors b », dont nous rechercherions la conformité de sa restriction avec les indices associés aux variables déjà envisagées en A.S.I..

Références

[1] Agrawal, R., Imielinsky, T., et Swami, A. (1993). Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD'93,* 207-216

[2] 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, 9-29

[3] Gras, R., Diday, E., Kuntz, P., et Couturier, R., (2001). Variables sur intervalles et variables-intervalles en analyse statistique implicative, *Actes du 8ème Congrès de la Société Francophone de Classification, Université des Antilles-Guyane,* 166-173

[4] Gras, R., et Kuntz, P.(2008). An overview of the Statistical Implicative, *Statistical Implicative Analysis*, R.Gras, E. Suzuki, F.Guillet and F.Spagnolo (Eds) Berlin Heidelberg: Springer-Verlag, 11-40.

[5] Gras, R., Régnier, J.C., et Guillet, F. (2009). *L'Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*, RNTI-E-16, Toulouse : Cépaduès Editions

[6] Guillet, F., et Hamilton, H.J. (Eds.) (2007). *Quality Measures in Data Mining*. New York : Springer-Verlag

[7] Hipp, J., Guntzer, U., & Nakhaeizadeh, J. (2000). Mining association rules: Deriving a superior algorithm by analyzing today's approach, *Proc. of 4th Eur. Conf. on Principles of Data Mining and Knowledge Discovery*, Lect. N. in Art. Int. 1910, 160-168.

[8] Lagrange, J.B., (1998). Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à modalités modales ordonnées. *Revue de Statistique Appliquée*, XLVI, 71-93.

[9] Lazlo, E. (1989). *La cohérence du réel : évolution, cœur du savoir.* Paris : Gauthier-Villars.

[10] Lenca, P., et Lallich, S. (2011). Le choix d'une bonne mesure de qualité, condition du succès d'un processus de fouille de données. Atelier Data Mining, Applications, Cas d'Etudes et Success Stories, *Extraction et Gestion des Connaissances*, *5-8.*

[11] Lenca, P., Vaillant, B., Meyer, P., et Lallich, S. (2007). Association rule interestingness measures: experimental and theoretical studies In F. Guillet et H.J. Hamilton (Eds.). *Quality Measures in Data Mining*. New York : Springer-Verlag 51-76

[12] Lerman, I.C. (1981). *Classification et analyse ordinale des données*. Paris : Dunod.

[13] Lerman, I.-C., et Guillaume, S. (2010). Analyse comparative d'indices d'implication discriminants fondés sur une échelle de probabilité. *Technical Report IRISA 1942/INRIA* 7187

[14] Métivier, M. (1968). *Notions fondamentales de la théorie des Probabilités*. Paris : Dunod Université.

[15] Muhlenbach, F., & Rakotomalala, R. (2005). Discretization of Continuous Attributes. In John Wang (Ed.). *Encyclopedia of Data Warehousing and Mining*, p. 397-402.

[16] Saporta, G. (2006. *Probabilités, Analyse de Données et Statistique*. Paris : Ed. Technip.

[17] Tomasis, A. (1977). *L'oreille et la vie*. Paris : R. Laffont.