Extraction de Règles en Incertain par la Méthode Implicative

Régis Gras*, Raphaël Couturier**, Fabrice Guillet*, Filippo Spagnolo***

** LINA– Ecole Polytechnique de l'Université de Nantes - BP 60601 44306 Nantes
regisgra@club-internet.fr et Fabrice.Guillet@polytech.univ-nantes.fr
** Institut Universitaire de Technologie de Belfort, BP 527, rue E. Gros, 90016 Belfort
Raphael.Couturier@iut-bm.univ.fcomte.fr
*** G.R.I.M., Department of Mathematics, University of Palermo
spagnolo@math.unipa.it.

Résumé. En relation avec des approches classiques de l'incertain, l'analyse statistique implicative (A.S.I.) peut apparaître innovante, particulièrement pour l'opérateur d'implication. L'article montre en effet que la notion de variables à valeurs intervalles et celle de variables-intervalles sont efficaces dans la détermination de leur distribution et dans la recherche de règles entre variables floues. De plus, elles apportent de riches informations sur la qualité de ces règles, tout en permettant d'étudier le rôle des variables supplémentaires dans l'existence de ces règles. Cette nouvelle perspective épistémologique de l'incertain ouvre d'intéressantes perspectives d'application.

1 Introduction

Partant du cadre défini et formalisé par (Lofti et al. 2001), (Dubois et al. 1987), ce texte vise à étudier les proximités formelle et sémantique des cadres de l'incertain et de l'analyse statistique implicative (A.S.I.) entre variables à valeurs intervalles et variables-intervalles (Gras et al. 2001). On s'intéresse particulièrement à l'opérateur « implication »avec lequel on extrait des règles d'association. Ce texte s'inscrit dans le cadre initié par (Gras 1979) sur l'analyse de données, A.S.I., qui vise à extraire et représenter, des règles d'association entre variables ou conjonctions de variables, du type $a \Rightarrow b$. Nous considérons celles qui croisent des sujets et des variables, présentant des modalités nettes ou floues. La qualité de la règle sera d'autant plus grande que son nombre de contre-exemples sera invraisemblablement petit sous l'hypothèse d'indépendance a priori, eu égard aux occurrences. .

2 problématique

Bien que les applications de la logique floue soient nombreuses en intelligence artificielle (par exemple en matière de diagnostic médical ou de reconnaissance des formes), plusieurs questions restent bien souvent latentes : comment obtient-on des distributions des degrés d'appartenance dans le cas de variables numériques ? Sur quelles connaissances sont-elles établies ? Sont-elles données a priori et mises à l'épreuve de la réalité ou bien sont-elles des construits ?S'il s'agit de ce dernier cas, quel processus d'extraction de connaissances à partir de données peut y conduire et quel type de règle peut-on alors extraire dans ce cadre ? Quelle signification peut-on donner à une règle associant deux sous-ensembles ou deux attributs flous ? On rejoint alors une des problématiques du data mining et de la qualité des règles.

3 Deux méthodes de construction de distributions floues par extraction de connaissances

Dans le cadre retenu, les distributions des degrés d'appartenance seront le fruit de l'interaction de connaissances objectives (vraie valeur de la variable, attribut net ou modificateur linguistique *consensuel*) et de connaissances subjectives. Dans la littérature, les degrés sont des données. D'où proviennent-elles ? Par ex., un échantillon d'individus étant donné on dispose **effectivement** de leur taille (un nombre) ou des caractères ou attributs nets :« petit », « moyen » et « grand » au vu d'une décision consensuelle du type : les caractères « petit », « moyen » et « grand » sont attribués **objectivement** au regard de leur taille mesurée. Face à ces données, on peut comparer le point de vue **subjectif** qui énonce :un sujet de 179 cm n'est pas petit, mais peut être considéré grand ou moyen. Plusieurs méthodes pour définir la distribution des attributs visent à effectuer un processus de « fuzzification » (Bernadet 2004).

3.1 Relation entre intervalles nets et attributs flous

Nous transformons les valeurs observées sur les sujets en sous-intervalles disjoints de variance inter-classe maximale afin de pouvoir attribuer à chaque sous-intervalle un attribut net de même désignation que celle attribuée aux attributs flous. Cette partition nette s'obtient par la méthode des nuées dynamiques. Enfin, pour chaque classe de similarité entre attribut net et attribut flou, nous déterminons le degré d'appartenance des sujets à une classe floue à partir de la mesure normalisée de typicalité associée à chaque individu. En effet, celle-ci, définie dans (Gras et al. 2001), rend compte d'un degré de responsabilité dans la proximité d'attributs, soulignant l'accord entre net et flou. Ainsi, nous disposerons d'une mesure vérifiant les axiomes de Zadeh relatifs au concept de « possibilité ». Mais, son avantage par rapport à la détermination subjective classique : elle s'établit à l'épreuve statistique de la réalité et elle varie avec la dilatation de l'ensemble des sujets. En résumé, les données initiales sont de 2 ordres : d'une part, des variables **objectives, consensuelles** aux valeurs numériques réparties sur des intervalles auxquels on associe respectivement un **attribut net** ; d'autre part, un **attribut flou** attribué **subjectivement** à chaque sujet.
Exemple : Sur 60 sujets de tailles T vraies variant de 168 et 198 cm, l'algorithme de la variance inter-classes maximale produit une hiérarchie entre les intervalles déterminés par : Tpeti de 168 à 174, Tmoy de 175 à 183, Tgran de 184 à 198. Les attributs flous sont notés respectivement TP, TM et TG.

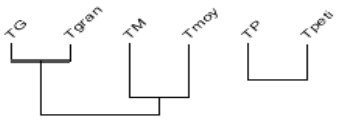

FIG. 1 – *Hiérarchie des similarités entre les intervalles*

On note que les attributs nets s'associent aux intervalles flous correspondants, ce que raisonnablement on pouvait attendre. Un sujet pourra donc posséder TM et TG suivant le point de vue du juge si sa taille n'est pas manifestement grande. Le logiciel CHIC (Couturier 2001) restitue les mesures de typicalité des sujets selon les 3 classes de similarité. Ces valeurs sont retenues comme degrés d'appartenance respectifs par rapport aux attributs flous.

3.2 Construction de l'histogramme d'une variable-intervalle à partir des données floues des sujets

La distribution des valeurs floues prises est donnée pour chaque sujet sur un intervalle. On cherche à en déduire une distribution de leurs degrés d'appartenance. L'objectif final est de définir une variable symbolique, variable-intervalle, qui soit l'histogramme d'un intervalle où on pourra déterminer des sous-intervalles optimaux selon le critère de la variance. Soit f_1, f_2, …,f_n les fonctions d'appartenance respectives des n sujets à un intervalle A. On suppose, par analogie avec les densités, que ces fonctions sont normalisées sur A. Dans la majorité des cas, chaque sujet contribue de la même façon à la densité, sinon une pondération adaptée y ramène. Alors la fonction $f=(f_1+f_2+ \ldots.+f_n)/n$ intègre en un histogramme la distribution des foncions d'appartenance. Il suffit ensuite de discrétiser A en une suite de points pondérés selon f ; puis d'appliquer sur A l'algorithme des nuées dynamiques pour obtenir une variable-intervalle dont on étudiera les relations implicatives avec les variables de ce type.

4 Règles d'association pour des variables numériques

On suppose maintenant que les distributions des variables floues sont connues selon 2 variables observées sur les mêmes sujets : taille et poids. On veut étudier, comme en ASI, les règles de déduction entre le prédicat taille et le prédicat poids, présentant des modalités, l'un **Taille** = {petit, moyen, grand}, l'autre **Poids** = {léger, moyen, lourd}. On dispose d'un tableau numérique donné des degrés d'appartenance aux modalités d'attributs flous d'un échantillon de 20 sujets. Les 3 premiers constituent *TAB 1 :* i_1, n'est donc pas très grand et pas très lourd, i_2 assez grand et assez lourd, i_3 plutôt grand et plutôt lourd.

	taille			**poids**		
	petit T_1	*moyen* T_2	*grand*T_3	*léger*P_1	*moyen*P_2	*lourd* P_3
i_1	8/15	5/15	2/15	7/14	4/14	3/14
i_2	1/14	6/14	7/14	2/15	5/15	8/15
i_3	0	7/16	9/16	1/16	6/16	9/16

TAB. 1 – *Valeurs prises par les modalités sur les 3 sujets.*

4.1 Un premier traitement de variables numériques

On effectue un traitement implicatif, selon l'A.S.I., en considérant les 6 variables tailles-poids comme variables numériques. On obtient le graphe implicatif (Régnier et al. 2004) à partir des sujets. Ainsi, les implications $T3 \Rightarrow P3$ et $P1 \Rightarrow T1$ sont valides à .90 et signifient que les règles grand $\Rightarrow$ lourd et léger $\Rightarrow$ petit, règle qui est sémantiquement contraposée de la première, sont acceptables. Une autre implication à un seuil >0.6 apparaît : $P_2 \Rightarrow T_1$.

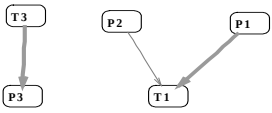

FIG. 2- *Graphe implicatif taille x poids*

Ces résultats ne s'opposent pas, bien entendu, au bon sens. Les autres règles d'association confirment une meilleure adéquation à la sémantique de l'implication qu'avec les approches de Reichenbach et Lukasiewicz. On ne retrouve pas, par ex. : léger $\Rightarrow$ grand. Mais, l'approche proposée ici présente l'inconvénient de considérer que les 6 modalités « taille » et « poids » sont actives dans le traitement et ne restituent pas, ainsi, les nuances de leur structure. Il est donc intéressant sémantiquement, de revenir à la considération de modalités de variables-intervalles qui apparaissent comme sous-intervalles d'une variable principale.

4.2 Second traitement par des variables à valeurs intervalles

Ce traitement (Gras et al. 2001) va permettre de prendre en compte de façon plus fine les nuances des observations prises selon des sous-ensembles flous et de répartir leurs valeurs de façon optimale sur un intervalle numérique [0 ;1], selon une partition dont l'utilisateur définit le nombre de classes. » pour chacun de 20 sujets. Nous disposons d'un nouveau *TAB 2* donnant les distributions des 6 modalités des 2 attributs « taille » et « poids » relativement à chacun des individus et les valeurs binaires prises par 2 variables supplémentaire »Femme », « Homme ». En voici les 2 premières lignes

	Taille petite t	**Taille moy. m**	**Taille grande T**	**Var supplé. Femme**	**Var. supplé. Homme**	**Poids léger L**	**Poids moy. o**	**Poids gran. P**
i1	0,7	0,4	0,3	1	0	0,8	0,3	0,1
i2	0,2	0,5	0,8	0	1	0,1	0,4	0,9

TAB. 2 – *Distributions des attributs flous « taille » et « poids »*

Par ex., le sujet i_1 admet un degré d'appartenance 0.7 à la classe des petits, 0.4 à celle des tailles moyennes et 0.3 à la classe des grandes tailles. De plus (variable supplémentaire) ce sujet est une femme et la distribution de ses degrés d'appartenance aux 3 classes de poids, sont resp. 0.8, 0.3 et 0.1. Le traitement emprunte cette fois la méthode des variables à valeurs intervalles. Comme dans le § 2, chaque modalité conduit à la construction de sous-intervalles optimaux, càd la détermination de sous-intervalles optimisant, du moins localement sinon globalement, l'inertie inter-classe. Avec CHIC pour le traitement de ce type de variable, on établit les règles telles que : si un sujet relève de l'intervalle t_i de la modalité t de l'attribut « taille » alors généralement il relève de l'intervalle p_j de la modalité p de l'attribut « poids ». Ainsi, si par ex., il a tendance à être plutôt petit, alors il a tendance à être plutôt léger.

Les partitions en 3 sous-intervalles calculées par CHIC sont données dans tableau 3.

tailles petites: t1 de 0 à 0.1 t2 de 0.2 à 0.5 t3 de 0.6 à 1	**tailles moyennes :** m1 de 0.1 à 0.3 m2 de 0.4 à 0.6 m3 de 0.8 à 0.8	**grandes tailles :** T1 de 0 à 0.1 T2 de 0.2 à 0.5 T3 de 0.8 à 0.9
poids légers : L1 de 0 à 0.2 L2 de 0.3 à 0.6 L3 de 0.8 à 1	**poids moyens :** o1 de 0.2 à 0.3 o2 de 0.4 à 0.5 o3 de 0.6 à 0.7	**poids lourds :** P1 de 0 à 0.1 P2 de 0.2 à 0.4 P3 de 0.7 à 0.9

TAB. 3 – *Partitions optimales calculées par CHIC*

Le graphe implicatif (*FIG 3*)au niveau de confiance 0.90 est également donné par CHIC :

On voit par exemple que l'individu de grande taille (T3) admet généralement un poids important (P3) et donc n'est pas considéré comme léger(L1). Ce sont les hommes qui apportent, et de très loin (risque de se tromper = 0.07), la plus importante contribution.

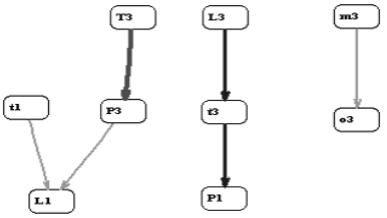

FIG. 3 –*Graphe implicatif taille x poids*

5 Conclusion

A l'aide de l'A.S.I., nous cherchons à objectiver la notion de degré d'appartenance. Situant le modèle d'implication entre attributs par rapport à des modèles classiques, nous mettons en évidence par un graphe, les relations implicatives entre modalités de variables numériques. Nous améliorons, semble-t-il, la formalisation de la sémantique en faisant référence à des variables-intervalles. Les règles les plus consistantes sont extraites selon leur qualité. Enfin, la relation entre variables extrinsèques et règles enrichissent notre connaissance de ces règles. Des situations réelles tenteront de valider cette approche de l'incertain.

Remerciements à Maurice Bernadet pour sa lecture du texte et ses précieux conseils

Références

Bernadet M.(2004), Qualité des règles et des opérateurs en découverte de connaissances floues. Mesure de qualité pour la fouille de données, Cepaduès, RNTI-E-1, pp 169-192

Couturier R. (2001),Traitement de l'analyse statistique implicative dans CHIC, Actes des Journées Fouille des données par l'analyse statistique implicative, IUFM Caen, pp 33-50.

Dubois D. et Prade H. (1987), Théorie des possibilités. Applications à la représentation des connaissances en informatique, Masson.

Gras R, (1979) Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'Etat, Rennes 1.

Gras R., Diday E., Kuntz P. et Couturier R. (2001), Variables sur intervalles et variables-intervalles en analyse implicative, 8ème Congrès SFC 17-21 décembre 2001, pp 166-173

Lotfi A. et Zadeh L.A. (2001), From computing with numbers to computing with words from manipulation of measurements to manipulation of perception, in Proceedings "Human and machine perception", Kluwer Academic, New York, 2001.