# Classification Conceptuelle avec Généralisation par Intervalles

Paula Brito*, Géraldine Polaillon**

*Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
mpbrito@fep.up.pt , http://www.fep.up.pt/docentes/mpbrito
**SUPELEC Science des Systèmes (E3S) - Département Informatique
Plateau de Moulon, 3 rue Joliot Curie, 91192 Gif-sur-Yvette cedex, France
geraldine.polaillon@supelec.fr

**Résumé.** Nous nous intéressons aux méthodes de classification hiérarchique ou pyramidale, où chaque classe formée correspond à un concept, i.e. une paire (extension, intension), considérant des données décrites par des variables quantitatives à valeurs réelles ou intervalles, ordinales et/ou prenant la forme de distribution de probabilités/fréquences sur un ensemble de catégories. Les concepts sont obtenus par une correspondance de Galois avec généralisation par intervalles, ce qui permet de traiter les données de différents types dans un cadre commun. Une mesure de la généralité d'un concept est alors calculée sous une forme commune pour les différents types de variables. Un exemple illustre la méthode proposée.

## 1 Introduction

La méthode de classification hiérarchique ou pyramidale proposée par Brito (voir, par exemple, Brito (1995)) permet de traiter des données multi-valuées où chaque classe formée est un concept, i.e., une paire (extension, intension). Cette méthode a été étendue par la suite à des données décrites par des variables modales, et permettant la prise en compte de l'existence de règles hiérarchiques entre variables catégoriques multi-valuées ou entre variables modales - voir Brito et De Carvalho (2008) pour une vision générale. Un critère numérique additionnel est défini, une mesure de "généralité", qui permet, à chaque étape, de choisir parmi les agrégations possibles. Les classes formées sont des "concepts", décrits en extension par la liste de ses membres, et en intension, par une expression conjonctive des valeurs prises par les variables constituant une condition nécessaire et suffisante d'appartenance à la classe. Les concepts sont obtenus grâce à la définition de correspondances de Galois pour chaque type de variables.

L'utilisation des treillis de Galois en analyse des données est d'abord due à Barbut et Monjardet (1970) et a été développée par Ganter et Wille (1999), d'abord pour des variables binaires ; cette approche a été appliquée à des variables non binaires sous condition préalable d'un recodage des données. Brito (1994) a défini des correspondances de Galois pour des variables quantitatives (réelles ou à valeurs intervalle) et qualitatives (mono ou multi-valuées), ce qui permet de traiter directement les données sans recodage ; par la suite cette approche a été étendue aux variables modales (Brito et Polaillon, 2005), i.e., variables qui prennent

comme valeurs des distributions de probabilités/fréquences sur un ensemble fini de classes ou de catégories (voir, par exemple, Noirhomme-Fraiture et Brito (2011)). Dans Polaillon et Brito (2011), nous proposons un cadre commun pour les variables quantitatives (réelles ou à valeurs intervalle), ordinales et modales, définissant un opérateur de généralisation qui calcule les intensions par intervalles de valeurs. Des variables de types différents peuvent ainsi être traitées directement ensemble. Pour les données de distribution, les concepts sont plus homogènes et plus facilement interprétables que ceux obtenus avec des opérateurs de généralisation par maximum et/ou minimum.

Dans (Polaillon, 2001) des méthodes sont présentées permettant d'obtenir une hiérarchie ou une pyramide par réduction du treillis de Galois des concepts. Dans ce papier nous proposons d'appliquer la correspondance de Galois définie dans Polaillon et Brito (2011) à la classification ascendante (hiérarchique ou pyramidale), dans la ligne des méthodes proposées précédemment - voir, par exemple, Brito (1995) et Brito et De Carvalho (2008).

## 2 La méthode de classification

Soit $E = \{\omega_1, ..., \omega_n\}$ l'ensemble de $n$ individus ou objets étant décrits par $p$ variables $Y_1, \ldots, Y_p$, $Y_j \rightarrow O_j$ et $O = O_1 \times \ldots \times O_p$ ; on appelle description un vecteur $d \in O$. Nous nous intéressons ici au cas où les variables $Y_j$ sont quantitatives (réelles ou à valeurs intervalle), ordinales ou modales.

Soit $(T, \leq_1)$ et $(W, \leq_2)$ deux ensembles ordonnés. Une correpondance de Galois est une paire d'applications $(f, g)$, $f : T \rightarrow W$ et $g : W \rightarrow T$, t.q. $f$ et $g$ sont antitones, et $h = gof$ et $h' = fog$ sont extensives ; $h$ et $h'$ sont alors des opérateurs de fermeture. L'application $f$ défini l'intension d'un ensemble $S \subseteq E$, alors que $g$ permet d'obtenir l'extension dans $E$ associée à une description $d \in O$. Un concept est une paire $(A, d)$ où $A \subseteq E, d \in O, A = g(d)$ et $d = f(A)$ ; $A$ est appelé l'*extension* du concept, et $d$ son *intension*.

La définition d'une méthode de classification où chaque classe formée doit correspondre à un concept repose alors sur la définition d'une correspondance de Galois appropriée. La méthode de classification conceptuelle ascendante proposée dans (Brito, 1995) peut être décrite par l'algorithme suivant : Soit $E = \{\omega_1, ..., \omega_n\}$ et $d^{(i)} = (d_1^{(i)}, \ldots, d_p^{(i)})$ la description associée à $\omega_i, i = 1, \ldots, n$. L'ensemble initial est l'ensemble des concepts, soit $\{(\omega_i, d^{(i)}), i = 1 \ldots, n\}$. Les classes sont alors construites récursivement : à chaque étape, une nouvelle classe $C$ est formée, par l'agrégation de classes préalables, $C_\alpha$ et $C_\beta$, $C = C_\alpha \cup C_\beta$ . Soit $d = f(C)$, alors les classes à agréger $C_\alpha$ et $C_\beta$ doivent vérifier les conditions suivantes : 1) $C_\alpha$ et $C_\beta$ peuvent être agrégées selon la structure de classification choisie (hiérarchie ou pyramide) ; 2) $g(d) = C$ i.e., aucun élément de $E$ qui n'appartient pas à $C$ appartient à l'extension de $d$ ; 3) La généralité de $d$ est minimale. Si aucune paire de classes $(C_\alpha, C_\beta)$ ne vérifie les conditions (1) et (2), l'algorithme effectue la réunion de plus de deux classes (adaptant les conditions d'agrégation). Le concept correspondant à une nouvelle classe formée est $(C, d) = (C, f(C))$ et chaque classe $C$ sera indéxée par $G(d) = G(f(C))$, la valeur de la mesure de généralité de $d = f(C)$. L'algorithme termine quand le concept $(E, f(E))$ est formé.

La généralité $G(d)$ d'une description $d \in O$ permet de définir un ordre total sur l'ensemble des descriptions $d$, ce qui revient à définir un ordre total sur l'ensemble des concepts. Pour des variables à valeurs ensemble, $G$ mesure la proportion de l'espace de description $O$ couverte par $d$ (voir (Brito, 1995)). Pour des variables à valeurs distributions, l'approche proposée (voir

Brito et De Carvalho (2008)) consiste à comparer la distribution donnée avec la distribution uniforme. Ici, on mesure la généralité $G(d)$ sous une forme commune pour les variables quantitatives, ordinales et modales, traitant ensemble des données décrites par ces différents types de variables.

## 2.1 La généralisation par intervalle

Dans le cas des variables numériques, ordinales et modales nous proposons un cadre unique de généralisation par intervalle, définissant l'application $f$ correspondante ; pour les données numériques on retrouve la généralisation par l'union (Brito, 1994). Soient $Y_1, \ldots, Y_p$ des variables réelles ou à valeur intervalle et $Y_j(\omega_i) = [l_{ij}, u_{ij}]$ (éventuellement $l_{ij} = u_{ij}$), et $A = \{\omega_1, \ldots, \omega_h\} \subseteq E$. La généralisation par l'union est définie par $f^U : P(E) \rightarrow I^p$, $I$ l'ensemble des intervalles de $\mathbb{R}$, avec l'ordre de l'inclusion, t.q. $f^U(A) = (I_1, \ldots, I_p)$, $I_j = [Min\{l_{ij}\}, Max\{u_{ij}\}]$, $\omega_i \in A$, $j = 1, \ldots, p$, i.e., $I_j$ est le plus petit intervalle qui contient toutes les valeurs prises par les éléments de $A$ pour $Y_j$. L'application $g^U : I^p \rightarrow P(E)$ avec $g^U((I_1, \ldots, I_p)) = \{\omega_i \in E : Y_j(\omega_i) \subseteq I_j, j = 1, \ldots, p\}$ donne l'extension d'une description. Le couple $(f^U, g^U)$ constitue une correspondance de Galois.

**Exemple 1 :** Soient 3 individus caractérisés par deux variables, nombre d'années de formation et salaire pendant les 5 dernières années (en milliers d'euros), représentés par $\omega_1 = (14, [1, 1.5])$, $\omega_2 = (17, [1.4, 1.8])$, $\omega_3 = (20, [1.60, 3])$. Soit $A = \{\omega_2, \omega_3\}$. La généralisation par l'union donne $f^U(A) = ([17, 20], [1.4, 3])$, qui décrit des individus dont le nombre d'années de formation varie entre 17 et 20 et dont le salaire des 5 dernières années a varié entre 1400 et 3000 euros ; $g^U(([17, 20], [1.4, 3])) = \{\omega_2, \omega_3\} = A$. Il en suit que $(A, ([17, 20], [1.4, 3]))$ est un concept pour la correspondance $(f^U, g^U)$.

Pour les variables modales, Brito et Polaillon (2005) définissent la généralisation par le maximum et par le minimum, obtenant ainsi deux correspondances de Galois distinctes ; ceci mène rapidement à une surgénéralisation, produisant des intensions $f(A)$, $A \subseteq E$, non informatives. Ici, nous effectuons la généralisation prenant pour chaque modalité l'intervalle de variation de sa probabilité/fréquence. Soit $Y_1, \ldots, Y_p$ des variables modales, avec $O_j = \left\{m_{j1}, \ldots, m_{jk_j}\right\}$ l'ensemble de modalités de la variable $Y_j$ ; pour la variable $Y_j$ et $\omega_i \in E$ on a $Y_j(\omega_i) = \left\{m_{j1}(p_{j1}^{(i)}), \ldots, m_{jk_j}(p_{jk_j}^{(i)})\right\}$, où $p_{jk_\ell}^{(i)}$ est la probabilité/fréquence associée à la modalité $m_{j\ell}$ de la variable $Y_j$ et l'individu $\omega_i$. Soit maintenant $M_j^I = \{m_{j\ell}(I_{j\ell}), \ell = 1, \ldots, k_j\}, m_{j\ell} \in O_j, I_{j\ell} \subseteq [0, 1]$ et $M^I = M_1^I \times \ldots \times M_p^I$. Pour $A \subseteq E$, la généralisation par intervalle est définie par $f^I : P(E) \rightarrow M^I$, $f^I(A) = (d_1, \ldots, d_p)$, avec $d_j = \{m_{j1}(I_{j1}), \ldots, m_{jk_j}(I_{jk_j})\}$, où $I_{j\ell} = \left[Min\{p_{j\ell}^{(i)}\}, Max\{p_{j\ell}^{(i)}\}\right], \omega_i \in A, \ell = 1, \ldots, k_j$, $j = 1, \ldots, p$. Soit $g^I : M^I \rightarrow E$ t.q. $g((d_1, \ldots, d_p)) =$
$\left\{\omega_i \in E : p_{j\ell}^{(i)} \in I_{j\ell}, \ell = 1, \ldots, k_j, j = 1, \ldots, p\right\}$. Le couple $(f^I, g^I)$ constitue une correspondance de Galois entre $(P(E), \subseteq)$ et $(M^I, \subseteq)$.

**Exemple 2 :** Soient 4 groupes d'individus décrits par tranche d'âge, a : âge $< 25$ ans, b : âge entre 25 et 60 ans, c : âge $> 60$ ans : $g_1 : (a(0.2), b(0.6), c(0.2))$ ; $g_2 : (a(0.3), b(0.3), c(0.4))$ ;

$g_3 : (a(0.1), b(0.6), c(0.3))$ ; $g_4 : (a(0.3), b(0.6), c(0.1))$ ; La généralisation par intervalle de $g_1$ et $g_2$ est donnée par $\{a\,[0.2, 0.3]\,, b\,[0.3, 0.6]\,, c\,[0.2, 0.4]\}$, décrivant des groupes où entre $20\%$ et $30\%$ des individus ont moins de 25 ans, entre $30\%$ et $60\%$ ont entre 25 et 60 ans, et entre $20\%$ et $40\%$ ont plus de 60 ans ; l'extension ne comprend que $g_1$ et $g_2$.

Le cas des variables ordinales a été traité par Pfaltz (2007), effectuant la généralisation par le maximum ou le minimum, le choix de l'opérateur étant fait individuellement pour chaque variable. Cependant, comme un de ces deux opérateurs est utilisé à chaque fois, la surgénéralisation n'est pas évitée. Nous proposons d'effectuer la généralisation dans le cas des variables ordinales considérant également un intervalle de valeurs.

**Exemple 3 :** Considérons 4 clients d'un magasin de vins ayant noté 2 vins, Vin 1, Vin 2 : $Client_1$ : (5, 5,) ; $Client_2$ : (5, 4,) ; $Client_3$ : (1, 2) ; $Client_4$ : (2, 1). La généralisation par intervalle des Clients 1 et 2 donne l'intension $([5, 5]\,, [4, 5])$, qui décrit des individus attribuant des notes élevées aux deux vins ; de même pour les Clients 3 et 4, $([1, 2]\,, [1, 2])$, décrit des individus attribuant des notes basses aux deux vins ; il s'en suit que $(\{\omega_1, \omega_2\}\,, ([5, 5]\,, [4, 5])$ et $(\{\omega_3, \omega_4\}\,, ([1, 2]\,, [1, 2])$ sont des concepts. Notons que l'intension par le maximum des Clients 1 et 2 est $(5, 5)$, décrivant des individus qui donnent n'importe quelle note aux deux vins ; l'extension correspondante comprend tous les quatre clients. D'autre part, l'intension par le minimum des Clients 3 et 4 est $(1, 1)$ décrivant également des individus qui attribuent n'importe quelle note aux deux vins, l'extension comprend tous les clients considérés. Dans les deux cas, on obtient des descriptions non-informatives, car sur-généralisées.

## 2.2 Évaluer la généralité

Une mesure de généralité permet de quantifier la généralité d'une description, et donc de choisir parmi les agrégations possibles à une étape donnée. Le principe sera que les classes associées à des concepts plus spécifiques doivent être d'abord formées. Ainsi, à chaque étape, parmi les classes qui peuvent être formées, on choisira de former celle dont l'intension du concept associée présente une moindre généralité.

On évalue la généralité d'une description $d = (d_1, \ldots, d_p)$ variable par variable, pour chaque variable $Y_j$ une valeur $G(d_j) \in [0, 1]$ est calculée, qui mesure la proportion de l'espace de description $O_j$ de $Y_j$ couverte par $d_j$. On mesure la généralité d'une description par la moyenne arithmétique des des valeurs obtenues pour chaque variable, $G(d) = \frac{1}{p}\sum_{j=1}^{p} G(d_j)$. La définition de $G(d_j)$ dépend du type de variable, c'est une mesure de l'ensemble couvert par $d_j$, et devra être croissante par rapport à l'inclusion. Pour des variables numériques $Y_j : E \to [L, U]$, telles que $d_j = [l_j, u_j]$, on a $G(d_j) = \dfrac{u_j - l_j}{U - L}$ ; la même définition s'applique aux variables ordinales. Pour les variables modales on a $d_j = \{m_{j1}(I_{j1}), \ldots, m_{jk_j}(I_{jk_j})\}$ avec $I_{j\ell} = [\underline{I}_{j\ell}), \overline{I}_{j\ell}]$ ; on définit alors $G(d_j) = \dfrac{1}{k_j}\displaystyle\sum_{\ell=1}^{k_j} \frac{\overline{I}_{j\ell} - \underline{I}_{j\ell}}{1 - 0} = \frac{1}{k_j}\sum_{\ell=1}^{k_j} (\overline{I}_{j\ell} - \underline{I}_{j\ell})$.

L'utilisation de la moyenne permet de donner la même importance aux différentes variables, et de comparer des descriptions de dimensions différentes. C'est le cas notamment pour les variables quantitatives ou ordinales, lorsqu'au moins un des intervalles $d_j = [l_j, u_j]$ est réduit à un point, i.e., lorsque $l_j = u_j$ pour au moins une valeur de $j \in \{1, \ldots, p\}$. La

mesure proposé précédement (voir Brito (1995)) évalue la généralité d'une description $d$ par la proportion de l'espace de description $O$ couverte par $d$, effectuant le produit des valeurs $G(d_j)$ ; cette définition suppose que les descriptions $d$ ont toutes la même dimension.

**Exemple 4 :** Considérons 2 groupes d'individus $g_1, g_2$ décrits par $Y_1$ = tranche d'âge, variable modale (voir Exemple 2) et $Y_2$ = revenu, à valeurs intervalle, $Y_2 : E \rightarrow [0, 10]$, $g_1 : (a(0.2), b(0.6), c(0.2), [2, 5])$; $g_2 : (a(0.3), b(0.3), c(0.4)), [1, 2.5])$. La description des 2 groupes est donnée par $d = (\{a\,[0.2, 0.3]\,, b\,[0.3, 0.6]\,, c\,[0.2, 0.4]\}, [1, 5])$, avec $G(d) = \frac{1}{2}\left[\frac{1}{3}\left((0.3 - 0.2) + (0.6 - 0.3) + (0.4 - 0.2)\right) + \frac{4}{10}\right] = \frac{1}{2}\left(\frac{0.6}{3} + 0.4\right) = 0.3$.

# 3 Application

Considérons 5 groupes d'individus décrits par $Y_1$ = tranche d'âge (voir Exemple 2) et $Y_2$ = revenu, $Y_2 : E \rightarrow [0, 10] : g_1 : (\{a(0.2), b(0.6), c(0.2)\}, [2, 5])$; $g_2 : (\{a(0.3), b(0.3), c(0.4)\}, [1, 2.5])$; $g_3 : (\{a(0.1), b(0.6), c(0.3)\}, [3, 6])$; $g_4 : (\{a(0.3), b(0.6), c(0.1)\}, [4, 8])$ et $g_5 : (\{a(0.5), b(0.3), c(0.2)\}, [1.5, 3])$. On construit une hiérarchie sur $E = \{g_1, g_2, g_3, g_4, g_5\}$. À la première étape, on peut former les classes $\{g_1, g_2\}, \{g_1, g_3\}, \{g_1, g_4\}, \{g_1, g_5\}, \{g_2, g_3\}, \{g_2, g_4\}, \{g_2, g_5\}, \{g_3, g_4\}, \{g_4, g_5\}$ qui sont des extensions d'un concept. La description de moindre généralité $d^{(6)} = (\{a([0.3, 0.5]), b([0.3, 0.3]), c([0.2, 0.4])\}, [1, 3])$ avec $G(d^{(6)}) = 0.167$ correspond à $C^{(6)} = \{g_2, g_5\}$. À l'étape suivante, on pourra former les classes $\{g_1, g_3\}$, $\{g_1, g_4\}$, $\{g_3, g_4\}, \{g_1, g_2, g_5\}$ et $\{g_2, g_4, g_5\}$ dont celle de moindre généralité est maintenant $C^{(7)} = \{g_1, g_3\}$, avec $d^{(7)} = (\{a([0.1, 0.2]), b([0.6, 0.6]), c([0.2, 0.3])\}, [2, 6])$ et $G(d^{(7)}) = 0.233$. À la troisième étape, les classes $\{g_1, g_3, g_4\}$, $\{g_2, g_4, g_5\}$ et $\{g_1, g_2, g_3, g_5\}$ correspondent à des concepts ; on forme la classe $C^{(8)} = \{g_1, g_3, g_4\}$, $d^{(8)} = (\{a([0.1, 0.2]), b([0.5, 0.6]), c([0.1, 0.3])\}, [2, 8])$ et $G(d^{(8)}) = 0.4$. Enfin, à la dernière étape, la classe $C^{(9)} \equiv E$ est formée, $d^{(9)} = (\{a([0.1, 0.5]), b([0.3, 0.6]), c([0.1, 0.4])\}, [1, 8])$, $G(d^{(9)}) = 0.5167$. La Figure 1 représente la hiérarchie indicée obtenue.

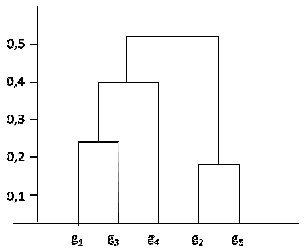


FIG. 1 – *Hiérarchie conceptuelle indicée*

# 4 Conclusion

Nous présentons une méthode de classification conceptuelle basée sur l'utilisation de correspondances de Galois récemment proposées pour la généralisation par intervalles de données

numériques, ordinales et modales et qui permet de fournir des descriptions plus homogènes, de manipuler ensemble plusieurs types de variables différents et limiter la sur-généralisation. Nous utilisons une mesure de généralité, qui sert de critère d'agrégation pour la classification ascendante, permettant de comparer des descriptions de différentes dimensions.

## Références

Barbut, M. et B. Monjardet (1970). *Ordre et Classification, Algèbre et Combinatoire, I & II*. Hachette.

Brito, P. (1994). Order structure of symbolic assertion objects. *IEEE Trans. on Knowledge and Data Engineering 6*(5), 830–835.

Brito, P. (1995). Symbolic objects : Order structure and pyramidal clustering. *Annals of Operations Research 55*, 277–297.

Brito, P. et F. A. T. De Carvalho (2008). Hierarchical and pyramidal clustering. In E. Diday et M. Noirhomme-Fraiture (Eds.), *Symbolic Data Analysis and the Sodas Software*, pp. 181–203. Wiley.

Brito, P. et G. Polaillon (2005). Structuring probabilistic data by Galois lattices. *MSH 169*(1), 77–104.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis, Mathematical Foundations*. Springer.

Noirhomme-Fraiture, M. et P. Brito (2011). Far beyond the classical data models : Symbolic data analysis. *Statistical Analysis and Data Mining 4*(2), 157–170.

Pfaltz, J. L. (2007). Representing numeric values in concept lattices. In J. Diatta et al. (Eds.), *Proc. 5th Int. Conf. Concept Lattices and their Applications*, pp. 260–269.

Polaillon, G. (2001). Pyramidal classification for interval data using Galois lattice reduction. In E. Diday et H.-H. Bock (Eds.), *Analysis of Symbolic Data*, pp. 433–440. Springer.

Polaillon, G. et P. Brito (2011). Homogénéité dans l'analyse conceptuelle : un cadre commun pour variables numériques, ordinales et modales. In *Actes SFC 2011, Orléans, France*.

## Summary

This paper deals with hierarchical or pyramidal conceptual clustering methods, where each formed cluster corresponds to a concept, i.e., a pair (extent, intent). We consider data presenting real or interval-valued numerical values, ordered values and/or probability/frequency distributions on a set of categories. Concepts are obtained by a Galois connection with generalisation by intervals, which allows dealing with different variable types on a common framework. In the case of distribution data, the obtained concepts are more homogeneous and more easily interpretable than those obtained by using the maximum and minimum operators previously proposed. A measure of generality of a concept is defined similarly for all these variable types. An example illustrates the proposed method.