Correspondances de Galois pour la manipulation de contextes flous multi-valués

Aurélie Bertaux[*,**], Florence Le Ber[*,***] et Agnès Braud[**]

[*]CEVH UMR MA 101 - ENGEES, 1 quai Koch, 67000 Strasbourg FRANCE
aurelie.bertaux, florence.leber@engees.u-strasbg.fr
http://engees-web.u-strasbg.fr/site/
[**]LSIIT UMR 7005, Bd Sébastien Brant, BP 10413, 67412 Illkirch cedex FRANCE
agnes.braud@urs.u-strasbg.fr
https://lsiit.u-strasbg.fr/fdbt-fr/index.php/Accueil
[***]LORIA UMR 7503, BP 35, 54506 Vandœuvre-lès-Nancy cedex FRANCE

Résumé. L'analyse formelle de concepts est une méthode fondée sur la correspondance de Galois et qui permet de construire des hiérarchies de concepts formels à partir de tableaux de données binaires. Cependant de nombreux problèmes réels abordés en fouille de données comportent des données plus complexes. Afin de traiter de tels problèmes, nous proposons une conversion de données floues multi-valuées en attributs histogrammes et une correspondance de Galois adaptée à ce format. Notre propos est illustré avec un jeu de données simples. Enfin, nous évaluons brièvement les résultats et les apports de cette correspondance de Galois par rapport à l'approche classique.

1 Introduction

L'analyse formelle de concepts est une méthode fondée sur la correspondance de Galois, qui permet de construire des hiérarchies de concepts formels à partir de tableaux de données binaires. Cependant de nombreux problèmes réels abordés en fouille de données comportent des données plus complexes, données multi-valuées ou floues. Pour prendre en compte de telles données, il faut donc étendre le modèle du treillis de Galois, comme cela a été proposé par (Messai et al., 2008; Bělohlávek et Vychodil, 2005; Stumme, 1999; Polaillon, 1998). Notre travail se situe dans cette lignée, et s'attache au traitement de données multi-valuées floues dans le cadre d'une application en hydrobiologie (Grac et al., 2006; Bertaux et al., 2007) que nous ne présentons pas ici par manque de place.

Après le rappel de quelques définitions, nous introduisons la notion de contexte flou multi-valué, puis présentons une transformation de tels contextes grâce à des attributs histogrammes et proposons des correspondances de Galois spécifiques pour les manipuler. Nous illustrons notre propos à l'aide d'un jeu de données simples et évaluons notre approche en comparaison à l'approche classique.

2 Définitions

Contextes binaires et treillis de Galois. Les treillis de Galois (Barbut et Monjardet, 1970; Davey et Priestley, 1990; Ganter et Wille, 1997) permettent de traiter des données binaires. On appelle *contexte* un triplet (O, T, I) où O est un ensemble d'objets, T un ensemble d'attributs, I une relation de $O \times T$ dans $\{0, 1\}$, et $I(o, t) = 1$ exprime que l'objet $o \in O$ possède l'attribut $t \in T$. Soient $X \subseteq O$ et $Y \subseteq T$. On définit $f : 2^O \rightarrow 2^T$ tel que $f(X) = \{y \in T | \forall x \in X : xIy\}$ est l'ensemble des attributs partagés par tous les objets de X. Par dualité, on définit $g : 2^T \rightarrow 2^O$ tel que $g(Y) = \{x \in O | \forall y \in Y : xIy\}$ est l'ensemble des objets qui possèdent tous les attributs de Y. Le couple (f, g) est appelé *correspondance de Galois* entre les ensembles O et T. Les opérations $h = g \circ f$ et $h' = f \circ g$ sont des opérateurs de fermeture, qui sont appelés *fermetures de Galois*. Un *concept* du treillis de Galois est un couple (X, Y), où l'*extension* X est telle que $X \subseteq O$ et $f(X) = Y$ et l'*intension* Y est telle que $Y \subseteq T$ et $g(Y) = X$.

Contextes flous multi-valués. Un contexte flou est un contexte (O, T, I) où la relation I est floue (Bělohlávek et Vychodil, 2005), c'est-à-dire qu'un objet peut posséder les attributs à des degrés divers. Un *contexte multi-valué* (O, T, M, I) se caractérise par un découpage des attributs en un ensemble de modalités M, un objet possédant au plus une modalité de chaque attribut (Ganter et Wille, 1997).

Nous introduisons la notion de *contexte flou multi-valué* comme une extension de ces deux notions, afin de formaliser des tableaux de données où les objets possèdent plusieurs modalités des attributs avec des degrés d'*affinité* variés. Un tel contexte peut s'écrire comme un quintuplet (O, T, M, A, I) où O est un ensemble d'objets, T un ensemble d'attributs, M un ensemble de modalités et A un ensemble d'affinités $\{0, 1, \ldots, \max_A\}$, avec $\max_A \geq 2$. I est une relation de $O \times T \times M$ vers A telle que $I(o, t, m) = a$ signifie que l'objet o possède la modalité m de l'attribut t avec une affinité a.

Attributs histogrammes. Pour manipuler des contextes flous multi-valués, nous avons choisi de convertir chaque triplet (attribut, modalités, affinités) en un *attribut histogramme* représentant la distribution des affinités selon les modalités de l'attribut d'origine. Considérons un contexte multi-valué flou tel que défini ci-dessus. À chaque attribut $t \in T$, possédant n_t modalités, est associé un ensemble H_t d'attributs histogrammes[1], notés h_t, tels que $h_t = \{a_1, \ldots, a_{n_t}\}$ où a_i prend valeur dans l'ensemble A. Soit $H = \cup_{t \in T} H_t$ l'ensemble de tous les attributs histogrammes. Le contexte multi-valué flou peut alors être réécrit comme un contexte binaire (O, H, I_H). Pour un objet $o \in O$ et un histogramme $h_t = \{a_1, \ldots, a_{n_t}\} \in H$, $I_H(o, h_t) = 1$ signifie que, dans l'ancien contexte, $I(o, t, m_i) = a_i$ pour tout $i \in [1, n_t]$.

Correspondances de Galois sur les attributs histogrammes. Le contexte (O, H, I_H) est un contexte binaire sur lequel peut être classiquement définie une correspondance de Galois. Toutefois, du fait du grand nombre d'attributs générés, les concepts obtenus sont peu nombreux et peu "fournis". Afin de pouvoir regrouper davantage d'objets dans un concept, nous utilisons une méthode proche de celles proposées par Polaillon (1998), méthodes qui associent à un ensemble d'objets symboliques l'union ou l'intersection de leurs valeurs d'attributs. Considérons

[1] Le cardinal de H_t est au plus $\max_A^{n_t}$.

alors la relation d'ordre suivante sur les attributs histogrammes :

$$(h_t^1 \leq h_t^2) \Leftrightarrow (a_1^1 \leq a_1^2) \wedge (a_2^1 \leq a_2^2) \wedge \ldots \wedge (a_{n_t}^1 \leq a_{n_t}^2)$$

où h_t^1 et h_t^2 sont deux histogrammes dérivant d'un même attribut t. Le minimum (respectivement le maximum) entre deux histogrammes h_t^1, h_t^2 est défini de la façon suivante :

$$\begin{array}{lcl} \min(h_t^1, h_t^2) & = & \{\min(a_1^1, a_1^2), \min(a_2^1, a_2^2), \ldots, \min(a_{n_t}^1, a_{n_t}^2)\} \\ \max(h_t^1, h_t^2) & = & \{\max(a_1^1, a_1^2), \max(a_2^1, a_2^2), \ldots, \max(a_{n_t}^1, a_{n_t}^2)\} \end{array}$$

Dans la suite on considère qu'un objet possède un histogramme pour chaque attribut d'origine, l'ensemble formant une séquence ordonnée. On note $\theta \in \Theta = \times_{t \in T} H_t$ cette séquence d'attributs histogrammes. $\theta(o)$ est la séquence d'attributs histogrammes de l'objet $o \in O$. La comparaison entre deux séquences θ_1 et θ_2 s'effectue simplement par comparaison deux à deux des histogrammes h_t^1, h_t^2 correspondant à chaque attribut t. Sur cette base, nous pouvons définir un couple (f_H, g_H) de fonctions sur les ensembles Θ et O :

$$\begin{array}{lcl} f_H : X \subseteq O & \rightarrow & f_H(X) = \{\theta \in \Theta | \min_{o \in X} \theta(o) \leq \theta \leq \max_{o \in X} \theta(o)\} \\ g_H : Y \subseteq \Theta & \rightarrow & g_H(Y) = \{o \in O | \min_{\theta \in Y} \theta \leq \theta(o) \leq \max_{\theta \in Y} \theta\} \end{array}$$

On peut montrer que le couple (f_H, g_H) constitue une correspondance de Galois, que nous appelerons correspondance min-max. Il s'agit d'une correspondance "floue" au sens où on ne regroupe pas des objets qui ont des attributs identiques, mais des objets qui ont des attributs *suffisamment* proches. Plus précisément, nous nous intéressons à déterminer des groupes d'objets compris entre des valeurs minimale et maximale d'attributs.

3 Illustration

Nous illustrons la notion de contexte flou multi-valué avec un exemple simple. Pierre, Paul et Jacques ont répondu à une enquête sur leurs habitudes matinales. Ils renseignent leurs préférences en indiquant par un '0' qu'ils ne prennent jamais telle boisson (modalités café, thé ou chocolat) ou telle tartine (modalités pain ou biscotte), par un '1' que cela leur arrive parfois et par un '2' que c'est fréquent. Le tableau 1 présente les données correspondantes.

	Boissons			Tartines	
	café	thé	chocolat	pain	biscottes
Pierre	0	1	2	0	2
Paul	2	0	2	2	0
Jacques	0	1	2	1	1

TAB. 1 – *Jeu de données concernant des habitudes de petit déjeuner.*

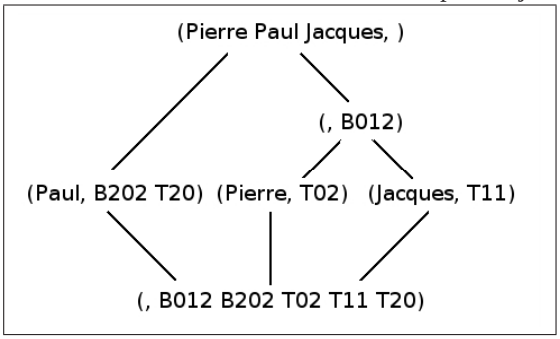
Conversion en histogrammes. Les attributs histogrammes correspondant aux triplets (attribut, modalités, affinités) de la table 1 sont constitués avec une lettre, 'B' pour Boissons et 'T' pour Tartines, à laquelle on accole des affinités en nombre et dans l'ordre correspondant aux modalités de l'attribut Boissons (3) ou Tartines (2). La relation binaire indique si Pierre, Paul

ou Jacques possède le profil décrit par l'histogramme. La table 2 présente le nouveau contexte binaire constitué avec les attributs histogrammes. Seuls les histogrammes utiles (i.e. associés à des individus) sont conservés. La figure1 présente le treillis obtenu par application d'une correspondance de Galois classique à ce jeu de données binaires[2].

Histogramme	B012	B202	T02	T11	T20
Pierre	1	0	1	0	0
Paul	0	1	0	0	1
Jacques	1	0	0	1	0

TAB. 2 – *Données histogrammes des habitudes de petit déjeuner.*

FIG. 1 – *Treillis de Galois des habitudes de petit déjeuner.*



Correspondance min-max. La correspondance min-max (f_H, g_H) permet de construire un treillis de Galois de la façon suivante (données du petit déjeuner). Initialement nous avons les couples $(o, \theta(o))$:

- A : **({Pierre}, {[B012 T02]})**
- B : **({Paul}, {[B202 T20]})**
- C : **({Jacques}, {[B012 T11]})**

On remarque que $f_H \circ g_H(\{\text{Pierre}\}) = g_H(\{[\text{B012 T02}]\}) = \{\text{Pierre}\}$. Chacun des couples A, B, C est fermé pour la correspondance min-max. Comparons maintenant ces couples deux à deux. Nous indiquons pour chaque comparaison, quels sont les histogrammes minimal et maximal issus de f_H et la fermeture qui en découle.

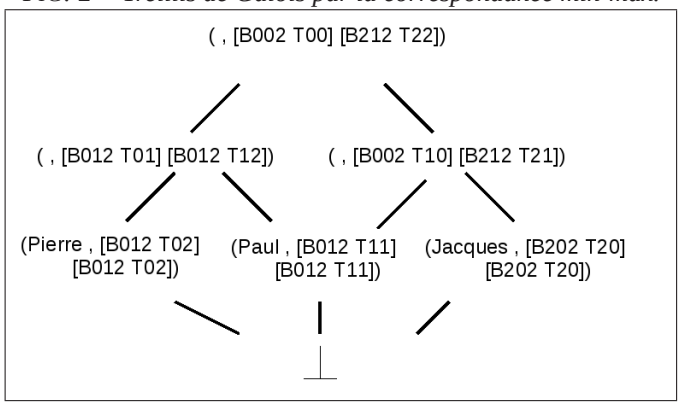
- A avec B :
  $f_H \circ g_H(\{\text{Pierre Paul}\}) = g_H(\{\theta | [\text{B002 T00}] \leq \theta \leq [\text{B212 T22}]\}) = \{\text{Pierre Paul Jacques}\}$
  On obtient le concept D : **({Pierre Paul Jacques}, {[B012 T02][B202 T20][B012 T11]})**.
- A avec C :
  $f_H \circ g_H(\{\text{Pierre, Jacques}\}) = g_H(\{\theta | [\text{B012 T01}] \leq \theta \leq [\text{B012 T12}]\}) = \{\text{Pierre Jacques}\}$
  On obtient le concept E : **({Pierre Jacques},{[B012 T02][B012 T11]})**.

[2]Les concepts d'un treillis héritent de manière ascendante de leurs attributs et de manière descendante de leurs objets. Ne sont affichés que les attributs et les objets propres à un concept et pas ceux hérités.

– B avec C :
$f_H \circ g_H(\{\text{Paul Jacques}\}) = g_H(\{\theta | [\text{B002 T10}] \leq \theta \leq [\text{B212 T21}]\}) = \{\text{Paul Jacques}\}$
On obtient le concept F : **({Paul Jacques},{[B202 T20][B012 T11]})**.

La figure 2 présente le treillis des concepts ainsi déterminés. En lieu et place de tous les attributs histogrammes d'un concept sont figurés leur minimum et leur maximum communs.

FIG. 2 – *Treillis de Galois par la correspondance min-max.*



Comparaison des treillis. Intéressons nous à la comparaison des deux treillis. La figure 1 montre que Pierre et Jacques partagent le plaisir de prendre du thé occasionnellement tout en préférant le chocolat : concept (Pierre Jacques, B012). Le concept E (Pierre Jacques, [B012 T01][B012 T12]) de la figure 2 est plus informatif car il nous renseigne non seulement sur les goûts de Pierre et Jacques en matière de boissons, mais nous précise qu'ils ont en commun de prendre des biscottes au petit déjeuner au moins de temps en temps (T01).

Soit maintenant Lucie, amatrice de chocolat et plus rarement de thé (B012). Si elle ne consomme que du pain (rarement ou souvent) (T10 ou T20) la correspondance min-max permet de l'associer à Paul et Jacques (concept F), alors que si elle ne consomme que des biscottes (rarement ou souvent) (T01 ou T02) Lucie serait associée à Pierre et Jacques (concept E).

Enfin la correspondance min-max sur les histogrammes permet de se rendre compte d'un point commun entre Pierre, Paul et Jacques grâce au concept D (Pierre Paul Jacques, [B002 T00][B212 T22]) : ils aiment tous beaucoup les boissons chocolatées (B002), ce qui n'apparaît pas de manière évidente avec la correspondance classique.

4 Conclusion

L'objectif de ce travail est de construire des concepts à partir de données multi-valuées floues. Nous avons pour cela considéré une conversion de ces données en histogrammes et une correspondance de Galois adaptée à ce format. L'intérêt de l'attribut histogramme est qu'il permet de représenter la répartition des affinités d'un objet pour les différentes modalités d'un attribut. L'intérêt de la correspondance de Galois min-max est de mettre en évidence les affinités minimales et maximales communes à un groupe d'objets. Ainsi il est possible de comparer

non seulement des distributions d'affinités sur les modalités mais également des "formes" de distribution comme par exemple B110 $\leq$ B210 $\leq$ B221, ce qui est particulièrement significatif pour des modalités ordonnées ou temporelles.

La correspondance min-max est en cours d'implantation et de test sur un jeu de données réelles concernant les caractéristiques de plantes aquatiques.

Remerciements. Les auteurs remercient l'Agence de l'Eau Rhin-Meuse et la Région Alsace pour leur soutien à ce projet de recherche.

Références

Barbut, M. et B. Monjardet (1970). *Ordre et classification – Algèbre et combinatoire*. Paris : Hachette.

Bertaux, A., A. Braud, et F. Le Ber (2007). Mining Complex Hydrobiological Data with Galois Lattices. In *Proc. of the 18th Int. Conference on Database and Expert Systems Application - Int. Workshop on Advances in Conceptual Knowledge Engineering (ACKE'07)*, pp. 519–523.

Bělohlávek, R. et V. Vychodil (2005). What is a fuzzy concept lattice. In *Concept lattices and Applications (CLA 2005)*, pp. 34–45.

Davey, B. et H. Priestley (1990). *Introduction to Lattices and Order*. Cambridge, UK : Cambridge University Press.

Ganter, B. et R. Wille (1997). *Formal Concept Analysis : Mathematical Foundations*. Springer Verlag.

Grac, C., A. Herrmann, F. Le Ber, M. Trémolières, A. Braud, A. Handja, et N. Lachiche (2006). Mining a database on alsatian rivers. In *Proc. of the 7th Int. Conference on Hydroinformatics (HIC 2006)*, Volume III, pp. 2263–2270.

Messai, N., M.-D. Devignes, A. Napoli, et M. Smaïl-Tabbone (2008). Many-valued concept lattices for conceptual clustering and information retrieval. In *Proc. of the 18th European Conference in Artificial Intelligence (ECAI 2008)*, pp. 127–131.

Polaillon, G. (1998). Organisation et interprétation par les treillis de galois de données de type multivalué, intervalle ou histogramme. Thèse de doctorat, Université Paris IX Dauphine.

Stumme, G. (1999). Hierarchies of conceptual scales. In *Proc. of Workshop on Knowledge Acquisition, Modeling and Management (KAW'99)*, Volume 2, pp. 78–95.

Summary

Formal concepts analysis is a method based on Galois connections, which builds hierarchies of formal concepts from binary data. But many real problems of data mining involve more complex data. To deal with such problems, we convert fuzzy many-valued contexts into histograms attributes and we use Galois connections suited to this format. Our approach is illustrated with a simple example. Finally, the results and the contributions of these connections are briefly evaluated with respect to the classical approach.