Représentation condensée de règles d'association multidimensionnelles

Alexandre Bazin*, Aurélie Bertaux** et Christophe Nicolle***

Le2i, Université de Bourgogne
21000 Dijon
*contact@alexandrebazin.com
**aurelie.bertaux@iut-dijon.u-bourgogne.fr
***cnicolle@u-bourgogne.fr

Résumé. La fouille de règles d'association est un problème qui a donné lieu à une littérature foisonnante, notamment dans les données binaires bidimensionnelles classiques. En particulier, la relation entre les ensembles fermés et les règles d'association est bien connue. Tel n'est pas le cas dans les données multidimensionnelles. Dans ce papier, nous montrons que la connaissance des n-ensembles fermés d'un tenseur booléen multidimensionnel est suffisante pour inférer la confiance de toutes les règles d'association multidimensionnelles.

1 Introduction

Le calcul de règles d'association (Agrawal et al. (1993)) est un problème important en fouille de données qui a donné lieu à une littérature foisonnante. Dès sa naissance, et pour pallier au grand nombre de motifs produits, l'accent a été mis sur la recherche d'ensembles réduits de règles contenant une information jugée intéressante. Comme souvent lorsque deux critères sont à optimiser – ici le nombre de règles et l'information contenue –, l'un d'eux prend le pas. Ainsi, dans le domaine des règles d'association, le nombre de règles est souvent vu comme primordial.

Le problème de la représentation de la totalité des règles – et donc de l'ensemble de l'information – est celui qui nous intéresse ici. Dans le cas de données binaires bidimensionnelles, le premier à être considéré, la question n'est plus ouverte. Nous savons que les règles peuvent être représentées de façon minimale par les ensembles fermés. Ce résultat, basé sur le fait que les fermés sont des représentants uniques de leurs classes d'équivalence vis à vis du support, a donné lieu à de nombreuses combinaisons avec les mesures d'intérêt utilisées pour réduire le nombre de règles au détriment de l'information.

Dans ce papier, nous considérons le cas des données binaires multidimensionnelles. Bien qu'il ait été moins étudié que le cas bidimensionnel, des moyens de réduire le nombre de règles ont déjà été proposées en généralisant la mesure d'intérêt la plus connue : la fréquence. Cependant, à notre connaissance, aucun résultat n'existe sur des représentations condensées de l'entièreté des règles. Nous nous proposons ici d'y remédier en montrant que, tout comme dans le cas bidimensionnel, les ensembles n-fermés d'une transformation d'un tenseur booléen

multidimensionnel sont suffisants pour dériver le support de toutes les associations et donc la confiance de toutes les règles.

Dans la Section 2, nous rappelons les définitions et propriétés connues et utiles des tenseurs, ensembles fermés et règles dans les tenseurs bi- et multidimensionnels. Dans la Section 3, nous montrons que le support naturel de toute association peut être calculé à partir du support d'associations particulières impliquant $n-1$ dimensions. Dans la Section 4, nous présentons une transformation du tensor permettant la dérivation du support d'associations par rapport à des ensembles de dimensions. Enfin, dans la Section 5, nous utilisons les résultats produits pour conclure que les n-ensembles fermés contiennent une information nécessaire et suffisante.

2 Définitions

2.1 Matrices, tenseurs et fermetures

Nous appelons *dimension* un ensemble $\mathcal{D}_i = \{d_1, \ldots, d_k\}$ d'éléments de même nature. Soient $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ un ensemble de dimensions et $\mathcal{R} \subseteq \prod_{\mathcal{D}_i \in \mathcal{D}} \mathcal{D}_i$ une *relation n-aire* entre les éléments des dimensions. Ensemble, $\mathcal{D}$ et $\mathcal{R}$ forment le *tenseur booléen* $\mathcal{T} = (\mathcal{D}, \mathcal{R})$, une matrice binaire n-dimensionnelle représentant des données. Ce tenseur est aussi appelé *contexte n-dimensionnel* ou *n-contexte* dans le domaine de l'analyse formelle de concepts. Le tenseur illustré dans la Figure 1 servira d'exemple tout au long de ce papier.

	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
c_1	×		×				×		
c_2		×		×	×	×	×		×
c_3			×		×			×	
	m_1			m_2			m_3		

FIG. 1 – *Tenseur tridimensionnel représentant des clients (c_1,c_2,c_3) achetant des produits (p_1,p_2,p_3) dans différents magasins (m_1,m_2,m_3).*

Dans $\mathcal{T}$, un *ensemble n-fermé* est un tuple $(X_1, \ldots, X_n)$ avec $X_i \subseteq \mathcal{D}_i$ tel que

$$\prod_{i \in \{1,\ldots,n\}} X_i \subseteq \mathcal{R}$$

et chaque composant est maximal pour cette propriété. Ainsi, dans le cas classique bidimensionnel, les 2-ensembles fermés sont les paires (A, B) dans lesquelles A et B sont maximaux tels que $\forall a \in A, \forall b \in B, (a, b) \in \mathcal{R}$. Dans ce cas, à la fois A et B sont dits *fermés*. Lorsque X est un sous-ensemble de l'une des deux dimensions, nous utiliserons $c(X)$ pour noter le plus petit (pour l'inclusion) ensemble fermé contenant X.

Dans notre exemple, nous trouvons les 3-ensembles fermés suivants :

(c_1,p_1p_3,m_1)	(c_1c_3,p_3,m_1)	(c_2,p_2,m_1m_2)	(c_1,p_1,m_1m_3)
(c_2,$p_1p_2p_3$,m_2)	(c_2c_3,p_2,m_2)	(c_2,p_1p_3,m_2m_3)	(c_1c_2,p_1,m_3)
(c_3,p_2,m_2m_3)	($\emptyset$,$p_1p_2p_3$,$m_1m_2m_3$)	($c_1c_2c_3$,$\emptyset$,$m_1m_2m_3$)	($c_1c_2c_3$,$p_1p_2p_3$,$\emptyset$)

L'ensemble des 2-ensembles fermés, ordonnés par la relation d'inclusion sur l'un de leurs deux composants, forme un treillis complet (Ganter et al. (1997)). De la même façon, l'ensemble des n-ensembles fermés, ordonnés par la relation d'inclusion sur l'un de leurs n-composants, forme un n-treillis complet (Voutsadakis (2002)). Les n-ensembles fermés peuvent être calculés grâce à DATA-PEELER (Cerf et al. (2008)).

2.2 Règles d'association dans le cas bidimensionnel

Traditionnellement, les règles d'association sont recherchées dans des relations binaires. Soient $\mathcal{D}_1$ et $\mathcal{D}_2$ deux dimensions, par exemple des *objets* et des *propriétés*, et $\mathcal{R} \subseteq \mathcal{D}_1 \times \mathcal{D}_2$ une relation binaire telle que $(o, p) \in \mathcal{R}$ signifie que l'objet o possède la propriété p.

	p_1	p_2	p_3	p_4	p_5
c_1	×	×			
c_2		×	×	×	
c_3		×		×	×
c_4	×		×		
c_5				×	×

FIG. 2 – *Tenseur bidimensionnel représentant des clients (c_1,c_2,c_3,c_4,c_5) achetant des produits (p_1,p_2,p_3,p_4,p_5).*

Soit A un sous-ensemble de $\mathcal{D}_2$. Le *support* de A, noté $s(A)$, est l'ensemble des éléments de $\mathcal{D}_1$ qui sont en relation avec tous les éléments de A. Une *règle d'association sur* $\mathcal{D}_2$ est un motif de la forme $A \xrightarrow{c} B$ dans lequel A et B sont des sous-ensembles de $\mathcal{D}_2$ et c, la *confiance* de la règle, est tel que

$$c = \frac{|s(A \cup B)|}{|s(A)|}$$

La règle représente le fait que $100c\%$ des éléments de $\mathcal{D}_1$ qui sont en relation avec tous les éléments de A sont aussi en relation avec tous les éléments de B. Nous utiliserons parfois les notations $A \rightarrow B$ pour la règle elle-même et $conf(A \rightarrow B)$ pour sa confiance.

Dans l'exemple de la Figure 2, la règle $p_1 \rightarrow p_2$ a une confiance de 0.5 car $s(\{p_1\}) = \{c_1, c_4\}$ et $s(\{p_1, p_2\}) = \{c_1\}$.

En l'absence de restrictions, il y a $2^{2|\mathcal{D}_2|}$ règles possibles. Il est donc nécessaire de n'en considérer qu'une partie ; idéalement la plus intéressante. Pour ce faire, un certain nombre de *mesures d'intérêt* ont été proposées (Zhang et al. (2009)). La première est la *fréquence*. Une règle $A \rightarrow B$ est dite *fréquente*, par rapport à un seuil $t \in [0, 1]$, si et seulement si $|s(A \cup B)| \geq t \times |\mathcal{D}_1|$. Ne s'intéresser qu'aux règles fréquentes permet ainsi de réduire significativement le nombre de motifs. Le seuil de fréquence peut être combiné à un seuil de confiance pour réduire encore le nombre de règles.

Cependant, le nombre d'ensembles fréquents peut se révéler toujours trop élevé. Pour y remédier, d'autres méthodes ont été proposées. Les propriétés suivantes font maintenant partie du folklore :

— $conf(A \rightarrow B) = conf(A \rightarrow A \cup B)$
— $conf(A \rightarrow B) = conf(c(A) \rightarrow c(B))$

La première indique que les règles importantes sont celles dont la conclusion contient la prémisse tandis que la seconde signifie qu'il suffit de ne considérer que les règles entre ensembles fermés.

Dans l'exemple de la Figure 2, la fermeture de $\{p_2\}$ est $\{p_2\}$ et celle de $\{p_2, p_5\}$ est $\{p_2, p_4, p_5\}$. De ce fait, les confiances de $p_2 \rightarrow p_2p_5$ et $p_2 \rightarrow p_2p_4p_5$ sont toutes deux égales à 1/3.

Cependant, construire des bases de règles d'association en utilisant toutes les règles entre ensembles fermés comparables n'est toujours pas suffisamment efficace puisqu'il peut y avoir jusqu'à $2^{|\mathcal{D}_2|}$ ensembles fermés. Afin de réduire encore plus le nombre de règles, il a été montré qu'il était suffisant de ne considérer que les règles de la forme $A \rightarrow B$ telles que $A = c(A)$, $B = c(B)$, $A \subset B$ et il n'y a aucun fermé X entre A et B. Cela correspond à une règle par arête dans le diagramme de Hasse du treillis des 2-ensembles fermés. La confiance de n'importe quelle règle peut alors être calculée en trouvant un chemin entre sa prémisse et sa conclusion dans le diagramme et en multipliant les confiances des règles parcourues.

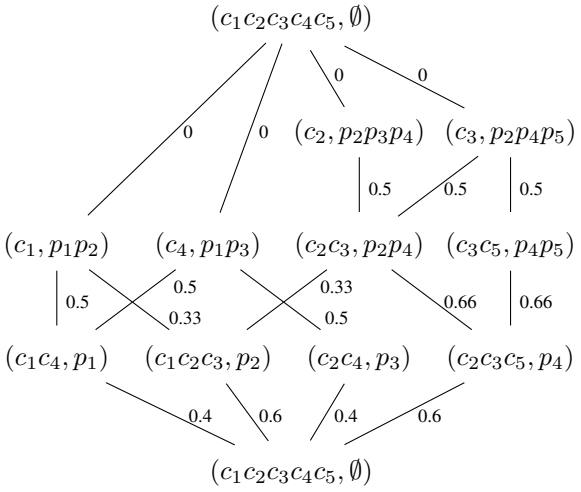

FIG. 3 – *Treillis des ensembles 2-fermés dans le tenseur bidimensionnel présenté dans la Figure 2 avec les confiances des règles d'association correspondantes aux arêtes.*

Sur le diagramme de Hasse illustré dans la Figure3, nous constatons que la confiance de $\emptyset \rightarrow p_4p_5$ peut être obtenue en multipliant les confiances de $\emptyset \rightarrow p_4$ et $p_4 \rightarrow p_4p_5$, ce qui nous donne $\frac{3}{5} \times \frac{2}{3} = 0.4$.

Luxenburger (Luxenburger (1991)) a montré que des ensembles de règles plus petit peuvent être obtenu en considérant uniquement un arbre couvrant du diagramme de Hasse. Cependant, l'utilisation de ces règles pour dériver des confiances implique de résoudre des problèmes d'optimisation linéaire, ce qui se révèle être trop chronophage pour la plupart des applications.

Puisque, par définition, les ensembles ont le même support que leur fermeture, les ensembles fréquents ont des fermetures fréquentes. Les deux approches pour réduire le nombre de règles peuvent donc être combinée en ne calculant que les règles entre ensembles fermés fréquents voisins (Lakhal et Stumme (2005)).

2.3 Règles d'association dans le cas multidimensionnel

Différentes généralisations des règles d'association dans les relations n-aires ont été étudiées. Dans Nguyen et al. (2011), les auteurs proposent ce qui est, pour nous, la plus générale. Nous la présentons ici et l'utilisons dans le reste de ce travail.

Soient $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ un ensemble de dimensions et $\mathcal{R} \subseteq \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ une relation n-aire entre les dimensions. Nous voulons extraire des "règles d'association" du tenseur $(\mathcal{D}, \mathcal{R})$. Cependant, contrairement au cas bidimensionnel, les motifs composant les règles peuvent impliquer différentes dimensions.

Soit $D \subseteq \mathcal{D}$ un ensemble de dimensions. Sans perte de généralité, nous supposons que $D = \{\mathcal{D}_1, \ldots, \mathcal{D}_{|D|}\}$. Soit $X_d \subseteq \mathcal{D}_d$, $\mathcal{D}_d \in D$, un ensemble non vide d'éléments de la dimension $\mathcal{D}_d$. L'ensemble de tuples $\prod_{\mathcal{D}_d \in D} X_d$ est appelé une *association sur* D et D est appelé son *domaine*. Nos règles d'association seront entre deux telles associations. Nous utiliserons $dom(X)$ pour noter le domaine d'une association X.

Nous omettrons les accolades des ensembles lorsque cela n'induit pas d'ambiguité et nous utiliserons la notation $X.Y$ à la place de $X \times Y$ pour représenter le produit cartésien de deux ensembles. Dans notre exemple, p_1 et $p_3.m_1$ sont des associations sur, respectivement, les domaines $\{\mathcal{D}_{produits}\}$ et $\{\mathcal{D}_{produits}, \mathcal{D}_{magasins}\}$ et $p_1 \rightarrow p_3.m_1$ est une règle d'association.

Soient $\mathcal{D}_i$ une dimension et $X = \prod_{\mathcal{D}_d \in Dom(X)} X_d$ une association. La *projection* $\pi_{\mathcal{D}_i}(X)$ de X sur $\mathcal{D}_i$ est égale à X_i si $\mathcal{D}_i \in Dom(X)$ ou à $\emptyset$ sinon.

Dans notre exemple, $\pi_{\mathcal{D}_{produits}}(p_3.m_1) = p_3$ $\pi_{\mathcal{D}_{clients}}(p_3.m_1) = \emptyset$ et $\pi_{\mathcal{D}_{magasins}}(p_3.m_1) = m_1$.

Dans le cas bidimensionnel, le support d'une association sur une dimension est un sous-ensemble de l'autre dimension. De la même façon, dans le cas multidimensionnel, le support d'une association est calculé sur les dimensions qui ne sont pas dans son domaine. Soit X une association, le *support de* X, noté $s(X)$, est l'ensemble $\{t \in \prod_{\mathcal{D}_i \in \overline{dom(X)}} \mathcal{D}_i \mid \forall x \in X, x.t \in \mathcal{R}\}$ des tuples dans le produit cartésien des dimensions absentes du domaine de X qui forment un élément de $\mathcal{R}$ avec un élément de X.

Dans notre exemple, nous avons que $s(p_1) = \{(c_1, m_1), (c_2, m_2), (c_1, m_3), (c_2, m_3)\}$ et $s(p_3.m_1) = \{c_1, c_3\}$.

Soient X et Y deux associations. Leur *union* est l'association $X \sqcup Y$ telle que, pour tout $\mathcal{D}_i \in \mathcal{D}$, $\pi_{\mathcal{D}_i}(X \sqcup Y) = \pi_{\mathcal{D}_i}(X) \cup \pi_{\mathcal{D}_i}(Y)$. Le motif $X \rightarrow Y$ est une *règle d'association multidimensionnelle sur le domaine* $dom(X \sqcup Y)$ si et seulement si $X \sqcup Y$ est une association sur $dom(X \sqcup Y)$.

Dans notre exemple, $p_1 \rightarrow p_3.m_1$ est une règle d'association sur le domaine $\{\mathcal{D}_{produits}, \mathcal{D}_{magasins}\}$.

Soit $X \rightarrow Y$ une règle sur $dom(X \sqcup Y)$. Si $dom(X)$ est différent de $dom(X \sqcup Y)$, les supports $s(X)$ et $s(X \sqcup Y)$ sont définis sur des ensembles différents et ne peuvent donc pas être comparés pour calculer la confiance de la règle. Le support de la prémisse doit donc être défini différemment. Le support de X par rapport à un domaine $D \supseteq dom(X)$ est défini par

$$s_{\overline{D}}(X) = \{t \in \prod_{\mathcal{D}_d \in \overline{D}} \mathcal{D}_d \mid \exists u \in \prod_{\mathcal{D}_i \in D \setminus dom(X)} \mathcal{D}_i \text{ such that } \forall x \in X, x.u.t \in \mathcal{R}\}$$

Grâce à ce support, nous pouvons définir la *confiance naturelle* de $X \rightarrow Y$ sur le domaine $D = dom(X \sqcup Y)$ par

$$conf(X \rightarrow Y) = \frac{|s(X \sqcup Y)|}{|s_{\overline{D}}(X)|}$$

Dans notre exemple, $s_{\mathcal{D}_{clients}}(p_1) = \{c_1, c_2\}$ et $s(p_1 p_3.m_1) = \{c_1\}$. Ainsi, la confiance de $p_1 \rightarrow p_3.m_1$ est

$$\frac{|\{c_1\}|}{|\{c_1, c_2\}|} = \frac{1}{2}$$

Ces règles d'association multidimensionnelles conservent la propriété que

$$conf(X \rightarrow Y) = conf(X \rightarrow X \sqcup Y)$$

Le nombre de ces règles est, évidemment, encore plus élevé que dans le cas bidimensionnel. Dans Nguyen et al. (2011), les auteurs utilisent la fréquence et la confiance pour le réduire. Il paraîtrait donc naturel d'imiter le cas bidimensionnel et de représenter aussi l'ensemble des règles d'associations n-dimensionnelles avec des n-ensembles fermés. Cependant, peu de résultats existent sur le sujet.

2.4 Transformations de tenseurs

Cette section présente les définitions de transformations de tenseurs que nous utilisons dans nos preuves.

Soient $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ un ensemble de dimensions, $\mathcal{R} \subseteq \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ une relation n-aire et $\mathcal{T} = (\mathcal{D}, \mathcal{R})$ un tenseur. Soient $D \subseteq \mathcal{D}$ un sous-ensemble des dimensions et $\mathcal{D}_d \in D$ une dimension. Le tenseur peut être transformé en :

— "fixant" des éléments d'une dimension
— combinant des dimensions

La première opération consiste en la restriction du tenseur à un sous-ensemble de l'une de ses dimensions. Soient $X_d \subseteq \mathcal{D}_d$ un ensemble d'éléments de la dimension $\mathcal{D}_d$ et $X_D = \{X_{j_1}, \ldots, X_{j_{|D|}}\}$ une collection d'ensembles d'éléments des dimensions dans D. Le tenseur $\mathcal{T}_{X_d} = (\mathcal{D} \setminus \mathcal{D}_d, \mathcal{R}_{X_d})$ avec

$$\mathcal{R}_{X_d} = \{(x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_n) \mid \forall x_d \in X_d, (x_1, \ldots, x_d, \ldots, x_n) \in \mathcal{R}\}$$

	p_1	p_2	p_3
c_1	×		×
c_2		×	
c_3			×

p_1	p_2	p_3
×		×

	m_1	m_2	m_3
c_1	×		
c_2		×	×
c_3			

FIG. 4 – *Transformations $\mathcal{T}_{m_1}$, $\mathcal{T}_{m_1,c_1}$ et $\mathcal{T}_{p_1,p_3}$ du tenseur $\mathcal{T}$ présenté dans la Figure 1.*

est construit en intersectant les "couches" de $\mathcal{T}$ correspondantes aux éléments de X_d. Le résultat est un tenseur $(n-1)$-dimensionnel. Lorsque plusieurs dimensions sont fixées simultanément, nous écrivons $\mathcal{T}_{X_D} = (((\mathcal{T}_{X_{j_1}})_{X_{j_2}}) \dots)_{X_{j_{|D|}}}$. La Figure 4 illustre cette transformation.

La seconde opération est le remplacement d'ensembles de dimensions par leur produit cartésien. Soient $\Omega = (\omega_1, \dots, \omega_m)$ une partition de $\mathcal{D}$ en m ensembles et

$$\mathcal{D}^\Omega = \{ \prod_{\mathcal{D}_i \in \omega_1} \mathcal{D}_i, \dots, \prod_{\mathcal{D}_j \in \omega_m} \mathcal{D}_j \}$$

le nouvel ensemble de dimensions. Le nouveau tenseur est alors $\mathcal{T}^\Omega = (\mathcal{D}^\Omega, \mathcal{R}^\Omega)$ avec

$$\mathcal{R}^\Omega = \{(x_1, \dots, x_m) \mid x_1.x_2.\dots.x_m \in \mathcal{R}\}$$

La Figure 5 illustre cette transformation.

	(p_1,m_1)	(p_2,m_1)	(p_3,m_1)	(p_1,m_2)	(p_2,m_2)	(p_3,m_2)	(p_1,m_3)	(p_2,m_3)	(p_3,m_3)
c_1	×		×				×		
c_2		×		×	×	×	×		×
c_3			×		×			×	

FIG. 5 – *Transformations $\mathcal{T}^{(\{\mathcal{D}_{clients}\},\{\mathcal{D}_{produits},\mathcal{D}_{magasins}\})}$ du tenseur $\mathcal{T}$ présenté dans la Figure 1.*

3 Dériver le support d'associations

Nous cherchons à identifier un ensemble restreint de règles d'association multidimensionnelles suffisantes pour dériver la confiance de toutes les autres. Pour ce faire, nous commencerons par montrer que la taille du support de n'importe quelle association intéressante peut être dérivée de la taille des supports de n-ensembles fermés. Nous supposons uniquement qu'une des dimensions n'apparait dans le domaine d'aucune association intéressante. Nous estimons cette supposition raisonnable car, en pratique, une dimension contient habituellement les "objets" ou "transactions" et ses éléments n'apparaissent pas dans les règles. Sans perte de généralité, nous supposerons que cette dimension est $\mathcal{D}_1$.

Tel que mis en évidence dans la Figure 6, le tenseur n-dimensionnel $\mathcal{T}$ peut être vu comme un empilement de tenseurs $(n-1)$-dimensionnels. De ce fait, la taille du support d'une association X est la somme des tailles de ses supports dans les différentes couches composant le tenseur.

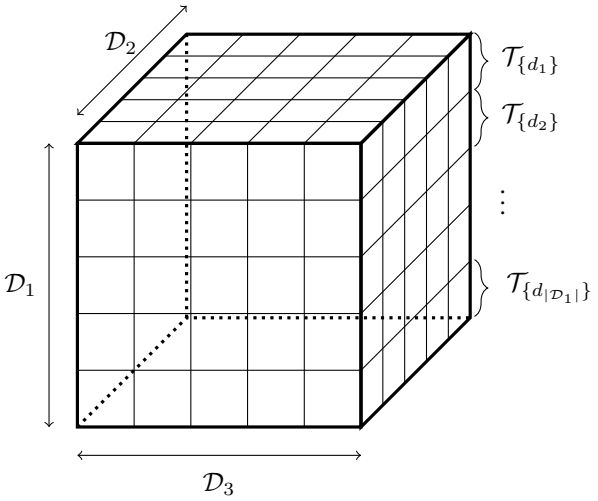

FIG. 6 – *Les tenseurs n-dimensionnels sont un empilement de tenseurs $(n-1)$-dimensionnels.*

Proposition 1. *Soit X une association. Soit*

$$D = \prod_{\mathcal{D}_i \in \mathcal{D} \setminus (dom(X) \cup \mathcal{D}_1)} \mathcal{D}_i$$

le produit cartésien de toutes les dimensions support de X à l'exception de $\mathcal{D}_1$. Nous avons que

$$|s(X)| = \sum_{d \in D} |s_d|$$

tel que s_d est le support de X dans $\mathcal{T}_d$..

Preuve. Par définition, tout $r \in \mathcal{R}_d$ est tel que $r.d \in \mathcal{R}$. Par conséquent, le support de X dans $\mathcal{T}_d$ est le sous-ensemble O de $\mathcal{D}_1$ tel que, $\forall o \in O$, $o.d$ est dans le support de X dans $\mathcal{T}$. Puisque les tuples $d \in D$ sont distincts deux-à-deux, la taille du support de X dans $\mathcal{T}$ est la somme des tailles des supports de X dans les différents $\mathcal{T}_d$. □

La Proposition 1 exprime le fait que les tailles des supports d'associations X et Y – et donc la confiance de la règle $X \rightarrow Y$ – dans un tenseur $\mathcal{T}$ peut se calculer à partir des tailles des supports de X et Y dans les différents tenseurs obtenus en fixant des éléments du produit cartésien de toutes les dimensions supports sauf $\mathcal{D}_1$. Dans ces tenseurs, $s(X)$ et $s(Y)$ sont tous deux sous-ensembles de $\mathcal{D}_1$.

Supposons que nous voulons connaître la taille des supports de p_1 et p_1p_3 dans notre exemple $\mathcal{T}$ et que $\mathcal{D}_{clients}$ est la dimension n'apparaissant pas dans les règles. Le produit cartésien de la seule dimension support qui n'est pas $\mathcal{D}_{clients}$ est $\{m_1, m_2, m_3\}$. Les supports de p_1 dans $\mathcal{T}_{\{m_1\}}$, $\mathcal{T}_{\{m_2\}}$ et $\mathcal{T}_{\{m_3\}}$ sont, respectivement, de taille 1, 1 et 2 donc $|s(p_1)| = 4$ dans $\mathcal{T}$. Les tailles des supports de p_1p_3 dans les mêmes tenseurs sont 1, 1 et 1 donc $|s(p_1p_3)| = 3$ dans $\mathcal{T}$. Ainsi, $p_1 \rightarrow p_1p_3$ a une confiance de $\frac{3}{4}$.

Proposition 2. *Soient X une association dans $\mathcal{T}$ et Z un élément du produit cartésien de dimensions supports de X. Le support de X dans $\mathcal{T}_Z$ est le support de $X \sqcup Z$ dans $\mathcal{T}$.*

Preuve. $\Rightarrow$. Soit S le support de X dans $\mathcal{T}_Z$. Par définition, $\forall s \in S$, $\forall x \in X$, $s.x \in \mathcal{R}_Z$. Puisque $\mathcal{T}_Z$ est construit à partir de $\mathcal{T}$ en intersectant les couches correspondantes à Z, nous avons que $s.x \in \mathcal{R}_Z$ implique que $s.Z.x \in \mathcal{R}$. Par conséquent, S est un sous-ensemble du support de $X \sqcup Z$ dans $\mathcal{T}$.

$\Leftarrow$. Soit S' le support de $X \sqcup Z$ dans $\mathcal{T}$. Par définition, $\forall s \in S$, $\forall x \in X$, $s.Z.x \in \mathcal{R}$. De la construction de $\mathcal{T}_Z$ nous obtenons que $s.Z.x \in \mathcal{R}$ implique que $s.x \in \mathcal{R}_Z$. Par conséquent, S' est un sous-ensemble du support de X dans $\mathcal{T}_Z$ et donc $S = S'$. □

Dans notre exemple, le support de $p_1.m_3$ dans $\mathcal{T}$ est $\{c_1, c_2\}$. Le support de m_3 dans $\mathcal{T}_{p_1}$ est aussi $\{c_1, c_2\}$.

La Proposition 2 implique que les supports des associations – et donc la confiance des règles – dans les tenseurs mentionnés dans la Proposition 1 peuvent être dérivés des supports des associations sur le domaine $\overline{\mathcal{D}_1}$ dans $\mathcal{T}$. Des Propositions 1 et 2, nous pouvons déduire que le support de n'importe quelle association dans $\mathcal{T}$ peut être dérivé des supports des associations sur le domaine $\overline{\mathcal{D}_1}$.

4 Règles entre associations sur différents domaines

Dans la Section 3, nous avons montré que la taille du support de n'importe quelle association peut être dérivée des tailles des supports d'associations sur $\overline{\mathcal{D}_1}$. Ce résultat est suffisant lorsque nous voulons calculer la confiance d'une règle entre deux associations sur le même domaine. Cependant, les règles de la forme $X \rightarrow Y$ avec $dom(X) \subset dom(X \sqcup Y)$ requièrent la connaissance du support $s_{\overline{dom(X \sqcup Y)}}(X)$ de X par rapport à $dom(X \sqcup Y)$. Dans cette section, nous montrons que le tenseur peut être transformé pour unifier les domaines des prémisses et conclusions de façon à ce que le support de n'importe quelle association par rapport à n'importe quel domaine puisse être dérivé d'associations sur $\overline{\mathcal{D}_1}$.

Comme nous l'avons présenté dans la Section 2.3, le support d'une association X par rapport à un domaine D est l'union des supports de toutes les associations pouvant être construites en augmentant de façon minimale X pour que D soit son domaine. Ce support peut être vu comme le support d'une association augmentée de façon à ce que chaque dimension additionnelle contienne un élément qui représente une disjonction sur l'entièreté de la dimension.

Définition 3. *Soit $\mathcal{T} = (\mathcal{D}_1, \ldots, \mathcal{D}_n, \mathcal{R})$ un tenseur. Nous définissons le tenseur $\mathcal{T}^\uparrow$ par $\mathcal{T}^\uparrow = (\mathcal{D}_1, \mathcal{D}_2 \cup \{\vee_2\}, \ldots, \mathcal{D}_n \cup \{\vee_n\}, \mathcal{R}^\uparrow)$ avec*

$$\mathcal{R}^\uparrow = \mathcal{R} \cup \{(x_1, \ldots, x_n) \mid \forall x_i = \vee_i, \exists x'_i \neq x_i \textit{ such that } (x_1, \ldots, x'_i, \ldots, x_n) \in \mathcal{R}^\uparrow\}$$

En d'autres termes, le tenseur $\mathcal{T}^\uparrow$ est construit à partir de $\mathcal{T}$ en ajoutant un élément à chaque dimension (à l'exception de $\mathcal{D}_1$) et en projetant les croix sur ces éléments jusqu'à saturation tel qu'illustré dans la Figure 7.

	p_1	p_2	p_3	$\vee_p$	p_1	p_2	p_3	$\vee_p$	p_1	p_2	p_3	$\vee_p$	p_1	p_2	p_3	$\vee_p$
c_1	×		×	×					×			×	×		×	×
c_2		×		×	×	×	×	×	×		×	×	×	×	×	×
c_3			×	×		×		×		×		×		×	×	×
	m_1				m_2				m_3				$\vee_m$			

FIG. 7 – *Le tenseur $\mathcal{T}^{\uparrow}$ correspondant à notre exemple $\mathcal{T}$ de la Figure 1. $\mathcal{D}_{clients}$ joue le rôle de $\mathcal{D}_1$.*

Définition 4. *Soient X une association dans $\mathcal{T}$ et $D \supseteq dom(X)$ un domaine. $X^{D\uparrow}$ est l'association dans $\mathcal{T}^{\uparrow}$ telle que*

$$\pi_{\mathcal{D}_i}(X^{D\uparrow}) = \begin{cases} \pi_{\mathcal{D}_i}(X) \cup \{\vee_i\} & \textit{si } \mathcal{D}_i \in D \\ \emptyset & \textit{sinon} \end{cases}$$

Proposition 5. *Soient X une association dans $\mathcal{T}$ et $D \supseteq dom(X)$ un domaine. $s_{\overline{D}}(X)$ dans $\mathcal{T}$ est égal à $s(X^{D\uparrow})$ dans $\mathcal{T}^{\uparrow}$.*

Preuve. Soit t un élément du support de $X^{D\uparrow}$ in $\mathcal{T}^{\uparrow}$. De la construction de $\mathcal{T}^{\uparrow}$, nous déduisons que, $\forall \mathcal{D}_i \in \overline{\mathcal{D}_1}$, $(x_1, \ldots, \vee_i, \ldots, x_n) \in \mathcal{R}^{\uparrow}$ implique que $(x_1, \ldots, x_i, \ldots, x_n) \in \mathcal{R}^{\uparrow}$ avec $x_i \neq \vee_i$. En suivant ce raisonnement récursivement sur les différentes dimensions dans D, nous obtenons que, pour tout $x \in X$, il existe un tuple $a \in \prod_{\mathcal{D}_i \in D \setminus dom(X)} \mathcal{D}_i$ tel que $t.a.x \in \mathcal{R}$. Ainsi, par définition, $t \in s_{\overline{D}}(X)$ dans $\mathcal{T}$. Par conséquent, $s(X^{D\uparrow})$ dans $\mathcal{T}^{\uparrow}$ est un sous-ensemble de $s_{\overline{D}}(X)$ dans $\mathcal{T}$.

Soit t' un élément de $s_{\overline{D}}(X)$ dans $\mathcal{T}$. Si t' n'est pas dans le support de $X^{D\uparrow}$, alors cela signifie qu'il n'y a aucun tuple $a \in \prod_{\mathcal{D}_i \in D \setminus dom(X)} \mathcal{D}_i$ tel que $t'.a.x \in \mathcal{R}$. Cela contredit notre supposition initiale. Par conséquent, $s_{\overline{D}}(X)$ dans $\mathcal{T}$ est un sous-ensemble de $s(X^{D\uparrow})$ dans $\mathcal{T}^{\uparrow}$ et ils sont donc égaux. □

La Proposition 5 implique que le support de X par rapport à un domaine D dans $\mathcal{T}$ est le même que le support de $X^{D\uparrow}$ dans $\mathcal{T}^{\uparrow}$. A partir de cela et des Propositions 1 et 2, nous pouvons déduire que la taille du support de n'importe quelle association X par rapport à n'importe quel domaine dans $\mathcal{T}$ peut être dérivé des tailles des supports d'associations sur $\overline{\mathcal{D}_1}$ dans $\mathcal{T}^{\uparrow}$.

Dans notre example $\mathcal{T}$ décrit dans la Figure 1, le support de m_3 par rapport au domaine $D = \{\mathcal{D}_{produits}, \mathcal{D}_{magasins}\}$ est $s_{\mathcal{D}_{clients}}(m_3) = \{c_1, c_2, c_3\}$. Dans le tenseur $\mathcal{T}^{\uparrow}$, l'association $m_3^{D\uparrow} = \vee_p.m_3 \vee_m$ a aussi $\{c_1, c_2, c_3\}$ pour support. décrit dans la Figure 7. De la même façon, le support de $p_1.m_3$ dans $\mathcal{T}$ et de $p_1.m_3^{D\uparrow} = p_1 \vee_p .m_3 \vee_m$ dans $\mathcal{T}^{\uparrow}$ est $\{c_1, c_2\}$.

5 Fermés et supports

Dans les Sections 3 et 4, nous avons montré que le support de n'importe quelle association par rapport à n'importe quel domaine peut être dérivé des supports des associations sur $\overline{\mathcal{D}_1}$ dans le tenseur $\mathcal{T}^{\uparrow}$. Il ne nous reste plus qu'à montrer que la connaissance de l'ensemble des n-ensembles fermés est suffisante pour retrouver ces supports.

Soit X une association sur le domaine $\overline{\mathcal{D}_1}$ dans $\mathcal{T}^{\uparrow}$. De par la définition d'une association et de son support, nous savons que $s(X).X \subseteq \mathcal{R}$. En d'autres termes, $s(X).X$ est une boîte de croix n-dimensionnelle dans le tenseur. Elle n'est pas nécessairement maximale sur toutes les dimensions mais le support lui-même l'est. De ce fait, il y a au moins un n-ensemble fermé $(s(X), C_2, \ldots, C_n)$ avec $\pi_{\mathcal{D}_i}(X) \subseteq C_i$, $\forall i \in \{2, \ldots, n\}$. Cela implique que $s(X) = s(\prod_{i \in \{2,\ldots,n\}} C_i)$. Dans le cas bidimensionnel, il n'y a qu'un tel 2-ensemble fermé pour X. Lorsque $n \geq 3$, il peut y en avoir plusieurs.

Définition 6. *Pour une association X, $c(X)$ désigne l'association résultante du produit cartésien des $n-1$ derniers composants d'**un** des n-ensembles fermés $(s(X), C_2, \ldots, C_n)$ avec $\pi_{\mathcal{D}_i}(X) \subseteq C_i$, $\forall i \in \{2, \ldots, n\}$.*

Par exemple, $c(\vee_p.m_1) = \vee_p.m_1 m_3 \vee_m$ parce que $s(\vee_p.m_1) = \{c_1c_2c_3\}$ et le triplet $(c_1c_2c_3, \vee_p, m_1m_3\vee_m)$ est un 3-ensemble fermé dans $\mathcal{T}^{\uparrow}$ (Figure 7).

Puisque X et $c(X)$ ont le même support, la connaissance de tous les n-ensembles fermés de $\mathcal{T}^{\uparrow}$ est suffisante pour dériver le support de n'importe quelle association X sur $\overline{\mathcal{D}_1}$ et donc de toute autre association, permettant ainsi de calculer la confiance de n'importe quelle règle d'association. Par extension, les règles de la forme $c(X) \rightarrow c(Y)$ telles que $c(X) \subseteq c(Y)$ sont suffisantes pour résumer toutes les autres règles.

6 Conclusion

Nous avons montré que le support des associations et donc la confiance des règles d'association dans un tenseur booléen multidimensionnel peuvent être dérivés de la connaissance des n-ensembles fermés d'une transformation du tenseur. Cela généralise les résultats connus dans le cas bidimensionnel puisque, dans ce contexte, la transformation ne modifie pas les 2-ensembles fermés.

Cependant, le fait que les supports des associations fréquentes puissent être dérivés des fermés fréquents, dans le cas bidimensionnel, ne se généralise pas au cas multidimensionnel. Les nombres d'associations fréquentes et de n-ensembles fermés mériteraient donc d'être à l'avenir comparés.

Remerciements

Ce projet a été partiellement financé par l'Union Européenne au travers du projet EUROSTAR PSDP. Les auteurs tiennent à remercier Nicolas Gros pour son apport.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD record*, Volume 22, pp. 207–216. ACM.

Cerf, L., J. Besson, C. Robardet, et J.-F. Boulicaut (2008). Data-peeler : Constraint-based closed pattern mining in n-ary relations. In *proceedings of the 2008 SIAM International conference on Data Mining*, pp. 37–48. SIAM.

Ganter, B., R. Wille, et C. Franzke (1997). Formal concept analysis : Mathematical foundations.

Lakhal, L. et G. Stumme (2005). Efficient mining of association rules based on formal concept analysis. In *Formal concept analysis*, pp. 180–195. Springer.

Luxenburger, M. (1991). Implications partielles dans un contexte. *Mathématiques, informatique et sciences humaines 29*(113), 35–55.

Nguyen, K.-N. T., L. Cerf, M. Plantevit, et J.-F. Boulicaut (2011). Multidimensional association rules in boolean tensors. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 570–581. SIAM.

Voutsadakis, G. (2002). Polyadic concept analysis. *Order 19*(3), 295–304.

Zhang, Y., L. Zhang, G. Nie, et Y. Shi (2009). A survey of interestingness measures for association rules. In *International Conference on Business Intelligence and Financial Engineering, 2009. BIFE'09.*, pp. 460–463. IEEE.

Summary

Association rules mining is a problem that gave rise to a rich literature, especially in classic binary bidimensional data. In particular, the relation between closed sets and association rules is well understood. This is not the case in multidimensional data. In this paper, we show that the knowledge of the closed n-sets of a multidimensional boolean tensor is enough to allow for the derivation of the confidence of every multidimensional association rule.