Extraction de motifs graduels clos

Sarra Ayouni*,** Sadok Ben Yahia*, Anne Laurent**, Pascal Poncelet**

*Faculté des Sciences de Tunis, 1060,Campus Universitaire, Tunis, Tunisie
sadok.benyahia@fst.rnu.tn
**LIRMM – CNRS, 161 rue Ada, Montpellier, France
{ayouni,laurent,poncelet}@lirmm.fr

Résumé. La découverte automatique de règles et motifs graduels ("plus l'âge d'une personne est élevé, plus son salaire est élevé") trouve de très nombreuses applications sur des bases de données réelles (e.g. biologie, flots de données de capteurs). Si des algorithmes de plus en plus efficaces sont proposés dans des articles récents, il n'en reste pas moins que ces méthodes génèrent un nombre de motifs tellement important que les experts peinent à les exploiter. Dans cet article, nous proposons donc une représentation condensée des motifs graduels en introduisant les concepts théoriques associés aux opérateurs de fermeture sur de tels motifs.

1 Introduction

Les outils de la fouille de données permettent d'extraire des motifs et des règles à partir des gros volumes de données. Récemment, l'extraction d'un nouveau type de motif a été étudiée : les motifs graduels (ou itemsets graduels), de la forme *Plus/Moins* X_1, ..., *Plus/Moins* X_n. De même que pour l'extraction d'itemsets classiques, on parle de motif graduel *fréquent* lorsqu'un nombre minimal de données de la base *supporte* le motif considéré. Plusieurs définitions du support ont été proposées dans la littérature. D'une manière générale, il s'agit de trouver à quel point les données de la base peuvent être ordonnées ou non pour que les relations de précédence présentes soient respectées (co-variations sur les attributs X_1, ..., X_n). Le support sert également à la définition d'algorithmes efficaces fonctionnant par niveau. Cependant ces algorithmes génèrent de très nombreux motifs et restent donc parfois difficilement utilisables par les experts. Face à ce problème, plusieurs méthodes sont envisageables : prise en compte de préférences utilisateurs, ordonnancement des motifs par rapport à des mesures de qualité, ou encore représentations condensées. Dans cet article, nous nous focalisons sur cette dernière méthode et introduisons la notion de motif graduel clos. D'une manière générale, nous appellons motif clos un motif qui n'a pas le même support que tout super-motif. Il concentre ainsi de l'information sur les sous-motifs et permet de définir une représentation condensée et sans perte (Boulicaut et Bykowski (2000); Pei et al. (2000); Pasquier et al. (1999)). L'ensemble des motifs fréquents peut alors être concentré en un sous-ensemble de motifs qui les représentent.

Dans cet article, nous utilisons la théorie de l'analyse formelle de concepts afin d'extraire de tels motifs graduels clos.

2 Travaux antérieurs

Dans ce que suit, nous présentons les approches qui ont été proposées pour l'extraction des motifs graduels à partir de bases de données et la terminologie utilisée. Pour les travaux sur L'analyse formelle de concepts (AFC), le lecteur est prié de se référer à (Ganter et Wille (1999)). Les motifs graduels sont extraits à partir de bases de schémas (X_1, ..., X_m) de domaines $dom(X_i)$ numériques (ou au moins munies d'un ordre total). Une base de données $\mathcal{D}$ est un ensemble de m-uplets de dom(X_1)×, ..., × dom(X_m). Dans ce contexte, un item et un itemset graduels sont définis comme suit :

Définition 1 ***Item/Itemset graduel*** *Soit un ensemble d'items $I \in \mathcal{I}$ tels que $i \in I$ est un item et $*$ est une variation avec $* \in \{\leq, \geq\}$. Un item graduel i^* est un item associé à une variation. Un itemset graduel est un ensemble d'items graduels noté (i_1^*, i_2^*, ..., i_n^*).*

Dans Hüllermeier (2002), les dépendances graduelles sont de la forme $A \rightharpoonup_t B$. Hüllermeier propose d'appliquer l'analyse par regression linéaire afin de les extraire et en calculer la qualité. Dans les deux approches Berzal et al. (2007) et Molina et al. (2007), les auteurs proposent de modifier la base initiale $\mathcal{D}$ en une autre $\mathcal{D}'$ contenant autant de lignes (ou tuples) que de couples d'objets distincts dans $\mathcal{D}$. Dans Laurent et al. (2009), les auteurs proposent, pour calculer le support, de considérer le *Kendall tau ranking correlation coefficient* qui calcule non pas la longueur du plus long chemin, mais le nombre de paires de n-uplets ordonnables dans la base de données pour être en accord avec le motif graduel considéré. Une autre définition du support d'un itemset graduel a été proposée dans (Di Jorio et al. (2008, 2009a), Di Jorio et al. (2009b)). Le support d'un itemset graduel $A_1^{*_1}, \ldots, A_p^{*_p}$, est défini par nombre maximal d'objets $\{r_1, \ldots, r_l\}$ pour lequel il existe une permutation π telle que $\forall\ j \in [1, l-1]$, $\forall\ k \in [1, p]$, nous avons $A_k(r_{\pi_j}) *_k A_k(r_{\pi_{j+1}})$.

Dans cet article, nous considérons l'approche de Di Jorio et al. (2009a) [1]. Étant donné $\mathcal{L}$ l'ensemble de tous les objets respectant la variation décrite par un itemset graduel, le support de cet itemset est alors défini comme :

Définition 2 *Soit s=$A_1^{*_1}, \ldots, A_p^{*_p}$ un itemset graduel, nous avons : supp(s) =* $\frac{\max_{L_i \in \mathcal{L}} |L_i|}{|\mathcal{D}|}$.

Les auteurs proposent une heuristique efficace en termes de performances, mais non exhaustive. Plus récemment Di Jorio et *al.* Di Jorio et al. (2009a), proposent une méthode efficace se basant sur les graphes de précédence. Dans cette méthode, appelée GRITE (GRadual ITemset Extraction), les données sont représentées dans un graphe dont les nœuds sont définis comme des objets de la base, et les liens expriment les relations de précédence dérivés à partir des attributs pris en compte. Malheureusement, dans toutes les approches que nous venons d'examiner, le problème de réduction de l'ensemble des motifs extraits n'est pas pris en considération. Nous proposons donc de définir une approche permettant de travailler sur une représentation condensée des motifs graduels fréquents. Notre approche s'appuie sur l'analyse formelle de concepts dont nous rappelons les principales caractéristiques ci-dessous.

1. Les auteurs remercient pour avoir fourni son implémentation de la recherche de motifs et règles graduels.

3 Opérateurs de Galois pour les motifs graduels

L'analyse formelle de concepts a été développée pour le traitement de données binaires. A notre connaissance, il n'existe pas dans la littérature de travaux traitant de l'application de cette théorie dans le contexte des motifs graduels. Nous proposons donc dans cet article d'étendre cette théorie à la gestion de la gradualité en formalisant un nouveau système de fermeture,Contrairement aux itemsets classiques dont la présence/absence peut être vérifiée pour chaque objet de la base de données, les variations d'un itemset graduel doivent être mesurées à partir d'un ensemble d'objets : peut-on ordonner les objets de la base pour qu'ils exhibent une augmentation simultanée sur les valeurs des attributs ? Cet ordonnancement est défini à partir d'une relation $\preceq$, et nous parlerons alors dans cet article de séquences d'objets, telles que définies ci-dessous.

3.1 Séquences d'objets

La recherche des motifs graduels est effectuée à partir d'une base de données d'objets. Nous définissons alors un ordre sur ces objets en fonction des motifs graduels considérés. Ainsi, nous considérons un ensemble d'objets (n-uplets) $\mathcal{O}$=$o_1, \ldots, o_n$ d'une base de données où chaque valeur o_i est définie sur un attribut dont le domaine est muni d'un ordre. Une séquence d'objets est une liste ordonnée de ces objets notée $\langle o_1, \ldots, o_m \rangle$.

Définition 3 *Soient S =* $\langle o_1, \ldots, o_p \rangle$ *et S'=* $\langle o'_1, \ldots, o'_m \rangle$ *deux séquences d'objets. S est* ***incluse*** *dans S' (*$S \subseteq S'$*) s'il existe des entiers* $1 \leq i_1 < i_2, \ldots, < i_p \leq m$ *tels que* $o_1 = o'_{i_1}, \ldots, o_p = o'_{i_p}$.

Définition 4 *Soit* $\mathcal{S}$ *un ensemble de séquences, une séquence d'objets* $S \in \mathcal{S}$ *est dite maximale si* $\nexists\ S' \in \mathcal{S}$, $S' \neq S$ *tel que* $S \subset S'$.

Définition 5 *L'****intersection*** *de deux séquences* S_1 *et* S_2 *est l'ensemble de sous-séquences maximales* $\mathcal{S}$ *telles que chaque séquence de* $\mathcal{S}$ *est une sous-séquence contenue à la fois dans* S_1 *et* S_2*, i.e.,* $S_1 \cap S_2 = \mathcal{S}$*, t.q.* $\forall\ s_i$ *de* $\mathcal{S}$, $s_i \subseteq S_1$ *et* $s_i \subseteq S_2$ *et* $\nexists\ s'_i \supset s_i$ *t.q.* $s'_i \subseteq S_1$ *et* $s'_i \subseteq S_2$.

Définition 6 *Soient* $\mathcal{S}$ *et* $\mathcal{S}'$ *deux ensembles de séquences d'objets.* $\mathcal{S}$ *est* ***inclus*** *dans* $\mathcal{S}'$ *(*$\mathcal{S} \preceq \mathcal{S}'$*) si* $\forall\ S \in \mathcal{S}$, $\exists\ S' \in \mathcal{S}'$ *t.q.,* $S \subseteq S'$.

Exemple 7 *Soient* S_1=$\langle o_1, o_2, o_4, o_7 \rangle$, S_2=$\langle o_2, o_5, o_1, o_4, o_6, o_8, o_7 \rangle$ *et* S_3=$\langle o_2, o_1, o_6, o_3, o_4 \rangle$ *trois séquences d'objets, nous avons* $S_3 \subseteq S_2$ *mais* $S_3 \nsubseteq S_1$ *et* $S_1 \cap S_2 = \{\langle o_2, o_4, o_7 \rangle, \langle o_1, o_4, o_7 \rangle\}$. *L'ensemble de séquences* $\mathcal{S}_1 = \{\langle o_5, o_6, o_7 \rangle, \langle o_2, o_4, o_7 \rangle\}$ *est inclus dans l'ensemble de séquences* $\mathcal{S}_2 = \{\langle o_5, o_6, o_8, o_7 \rangle, \langle o_1, o_2, o_4, o_7 \rangle\}$.

Proposition 1 *Soit* $\mathcal{S}$ *un ensemble de séquences maximales,* $\preceq$ *définit un ordre partiel sur* $\mathcal{P}(\mathcal{S})$.

3.2 Opérateurs de fermeture graduels

Dans cet article, nous proposons une nouvelle définition de la connexion de Galois graduelle qui prend en compte les itemsets graduels.

Définition 8 *Un contexte formel graduel est défini comme un quadruplet $\mathcal{K}$= ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{Q}$, R) décrivant un ensemble d'objets $\mathcal{O}$, un ensemble fini $\mathcal{I}$ d'attributs (ou items), un ensemble fini de valeurs quantitatives $\mathcal{Q}$ et une relation binaire $\mathcal{R}$ (i.e., $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$). Chaque couple $(o, i^q) \in \mathcal{R}$ correspond au fait que la valeur de l'attribut (item) i appartenant à $\mathcal{I}$ pour l'objet o appartenant à $\mathcal{O}$ est q appartenant à $\mathcal{Q}$.*

Soit $\mathcal{K}$= ($\mathcal{O}, \mathcal{I}, \mathcal{Q}, \mathcal{R}$) un contexte formel graduel, nous définissons ci-dessous les deux opérateurs f et g :

$$f : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{I})$$
$$f(S) = \{i^* | \forall\, s \in S, \forall o_l, o_k \in s \text{ t.q. } (o_l, i^{q_1}), (o_k, i^{q_2}) \in \mathcal{R} \text{ et } k < l \text{ nous avons } q_1 * q_2\}$$

La fonction f retourne tous les items graduels qui respectent toutes les séquences de S, ainsi que leurs variations respectives.

$$g : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{S})$$
$$g(I) = \{s \in S | \ s \text{ est maximale dans } S \text{ et } \forall o_l, o_k \in s \text{ t.q. } k < l \ et \ (o_l, i^{q_1}), (o_k, i^{q_2}) \in \mathcal{R}, \forall i^* \in I \text{ nous avons } q_1 * q_2 \ \}$$

La fonction g retourne l'ensemble des séquences maximales respectant les variations de tous les items de I.

Les deux fonctions g et f sont définies respectivement sur l'ensemble des parties de $\mathcal{I}$ et sur l'ensemble des parties des séquences de $\mathcal{S}$. Étant donné le fait que l'intersection d'un ensemble de séquences d'objets peut résulter en plus d'une séquence, nous considérons l'ensemble des parties de séquences. La fonction f est appliquée sur un ensemble de séquences tandis que g s'applique sur un ensemble d'attributs graduels. L'ensemble des itemsets graduels peut être ordonné par la relation d'inclusion ensembliste classique $\subseteq$ tandis que l'ensemble des séquences est ordonné par la relation $\preceq$. À partir des définitions et propositions introduites ci-dessus, nous pouvons démontrer que nous construisons un contexte permettant l'utilisation de la connexion de Galois pour le cas graduel.

Proposition 2 *Pour les ensembles de séquences S et $S' \in \mathcal{S}$, et les ensembles d'itemsets graduels I et $I' \in \mathcal{I}$ les propriétés suivantes sont vérifiées :*

1) $S \preceq S' \Rightarrow f(S') \subseteq f(S)$ **1')** $I \subseteq I' \Rightarrow g(I') \subseteq g(I)$

2) $S \preceq g(f(S))$ **2')** $I \subseteq f(g(I))$

Proposition 3 *Les fonctions composites $f \circ g$ et $g \circ f$ forment deux opérateurs de fermeture, définies respectivement sur les ensembles de séquences et l'ensemble des itemsets.*

Ces propositions permettent de considérer ce contexte pour la définition et l'extraction de représentations condensées de motifs graduels, comme défini ci-dessous.

Définition 9 ***Concept graduel formel*** *Le couple (S, I), tel que $S \in \mathcal{S}$ et $I \in \mathcal{I}$, est un concept graduel si $f(S) = I$ et $g(I) = S$. O est l'extension et I l'intension du concept graduel.*

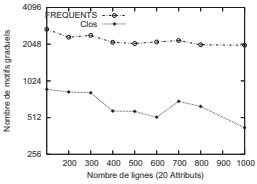
Définition 10 ***Itemset graduel clos*** *Soit le contexte formel $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{Q}, \mathcal{R})$, un sous-ensemble graduel $I \subseteq \mathcal{I}$ est un itemset graduel clos s'il est égal à sa fermeture, i.e., $f \circ g(I) = I$.*

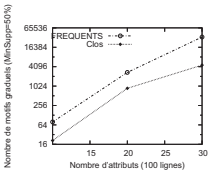
Définition 11 ***Générateur graduel minimal*** *Un itemset graduel $h \subseteq \mathcal{I}$ est dit générateur graduel minimal d'un itemset graduel clos I si $f \circ g(h) = I$ et il n'existe pas $h' \subseteq \mathcal{I}$ tel que $h' \subset h$. L'ensemble $\mathcal{GGM}$ des générateurs graduels minimaux d'un itemset graduel clos I est défini comme suit : $\mathcal{GGM} = \{ h \subseteq \mathcal{I} \mid f \circ g(h) = I \wedge \nexists\, h' \subset h$ tel que $f \circ g(h') = I \}$*

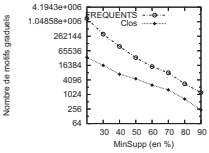
Proposition 4 *L'ensemble de concepts graduels formels $\mathcal{GC}_{\mathcal{K}}$ extraits du contexte $\mathcal{K}$, ordonnés selon l'inclusion ensembliste, forme un treillis complet $\mathcal{L}_{\mathcal{K}} = (\mathcal{GC}_{\mathcal{K}}, \subseteq)$, que nous nommons treillis de Galois graduel.*

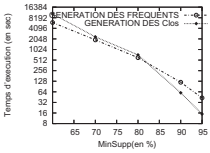
4 Expérimentations

Afin de montrer la concision apportée par notre méthode, nous comparons le nombre de motifs clos extraits par rapport au nombre de motifs graduels fréquents. Les expérimentations sont menées sur des jeux de données synthétiques générés à l'aide d'une version modifiée de IBM Synthetic Data Generation Code for Associations and Sequential Patterns [2]. Notons que ces bases sont très denses et produisent, même pour des valeurs de support minimum élevées, un très grand nombre de motifs fréquents. Dans le cadre de nos expérimentations, nous nous sommes intéressés à la variation du nombre de motifs clos par rapport au nombre de motifs fréquents en fonction de la valeur de support minimum (minsupp), du nombre de lignes et d'attributs de la base de données, ainsi qu'au temps de calcul. Dans cet article, nous visons









en effet à valider l'importance de la réduction du nombre de motifs extraits. Comme décrit ci-dessus, l'approche est un *post-traitement* de Di Jorio et al. (2009a). Les temps de calcul sont donc un peu plus longs, comme le montre la Figure 4. Cependant, une version intégrée de recherche des clos est en cours d'implémentation et permettra d'extraire les clos en des temps considérablement réduits.

5 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés aux représentations condensées d'ensembles de motifs graduels de la forme *Plus/Moins* X_1, ..., *Plus/Moins* X_n. Nous définissons les opé-

2. www.almaden.ibm.com/software/projects/hdb/resources.shtml

rateurs nécessaires, et montrons l'intérêt de notre approche à travers diverses expérimentations. Les perspectives associées à ce travail sont nombreuses, notamment pour intégrer le calcul des clos au long du processus et non en post-traitement comme effectué actuellement et pour définir les opérateurs de fermeture dans le cas de motifs graduels flous et de séquences graduelles.

Références

Berzal, F., J. Cubero, D. Sánchez, M. Vila, et J. Serrano (2007). An alternative approach to discover gradual dependencies. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15*(5), 559– 570.

Boulicaut, J.-F. et A. Bykowski (2000). Frequent closures as a concise representation for binary data mining. In *PADKK '00 : Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, London, UK, pp. 62–73. Springer-Verlag.

Di Jorio, L., A. Laurent, et M. Teisseire (2008). Fast extraction of gradual association rules : a heuristic based method. In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology (CSTST '08)*, pp. 205–210.

Di Jorio, L., A. Laurent, et M. Teisseire (2009a). Extraction efficace de règles graduelles. In *EGC'09*, pp. 199–204.

Di Jorio, L., A. Laurent, et M. Teisseire (2009b). Mining frequent gradual itemsets from large databases. In *Proceedings of the International Conference Intelligent Data Analysis (IDA'09)*. LNCS, Springer Verlag.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis*. Springer-Verlag.

Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. In *Proceedings of the International Conference PKDD'2002, Helsinki, Finland*, pp. 200–211.

Laurent, A., M.-J. Lesot, et M. Rifqi (2009). Graank : Exploiting rank correlations for extracting gradual dependencies. In *Proc. of FQAS'09*.

Molina, C., J. Serrano, D. Sánchez, et M. Vila (2007). Measuring variation strength in gradual dependencies. In *Proceedings of the International Conference EUSFLAT'2007, Ostrava, Czech Republic*, pp. 337–344.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Discovering frequent closed itemsets for association rules. In *ICDT '99 : Proceedings of the 7th International Conference on Database Theory*, London, UK, pp. 398–416. Springer-Verlag.

Pei, J., J. Han, et R. Mao (2000). Closet : An efficient algorithm for mining frequent closed itemsets. In *Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

Summary

Mining gradual patterns and rules (like "The older, the higher the salary") is more and more studied as it has numerous applications for real world databases (e.g. biology, stream mining). Algorithms have recently been designed to mine for such patterns. However, they suffer from the fact that the number of patterns extracted by the methods is too huge to be easily managed by the end-users. We thus propose here to define condensed representations of gradual patterns by defining closure operators, and we show the interest of our method with experiments.