Extraction efficace de règles graduelles

Lisa Di Jorio*, Anne Laurent* Maguelonne Teisseire*

*LIRMM – Université de Montpellier 2 – CNRS
161 rue Ada, 34392 Montpellier – FRANCE
{dijorio, laurent, teisseire}@lirmm.fr
http://www.lirmm.fr/~{dijorio, laurent, teisseir}

Résumé. Les règles graduelles suscitent depuis quelques années un intérêt croissant. De telles règles, de la forme "*Plus (moins)* A_1 *et ... plus (moins)* A_n *alors plus (moins)* B_1 *et ... plus (moins)* B_n" trouvent application dans de nombreux domaines tels que la bioinformatique, les contrôleurs flous, les relevés de capteurs ou encore les flots de données. Ces bases, souvent composées d'un grand nombre d'attributs, restent un verrou pour l'extraction automatique de connaissances, car elles rendent inefficaces les techniques de fouille habituelles (règles d'association, clustering...). Dans cet article, nous proposons un algorithme efficace d'extraction d'itemset graduels basé sur l'utilisation des treillis. Nous définissons formellement les notions de gradualité, ainsi que les algorithmes associés. Des expérimentations menées sur jeux de données synthétiques et réels montrent l'intérêt de notre méthode.

1 Introduction

L'évolution des capteurs, de plus en plus précis, robustes et abordables, permet l'acquisition de nombreuses mesures fiables, produisant ainsi des bases de données numériques denses et volumineuses. Cependant, même si la fouille de données quantitatives est un domaine étudié depuis plusieurs années (Srikant et Agrawal (1996)), la densité des bases pose de nouvelles problématiques, car elle rend inefficaces les techniques de fouille habituelles. Pourtant, les experts sont de plus en plus en attente de méthodes efficaces afin de prendre des décisions ou d'analyser différents comportements. Beaucoup de domaines sont concernés, comme par exemple le domaine biomédical, où les principales découvertes passent par l'analyse du génome (contenant plusieurs milliers de gènes, pour peu de patients), ou encore le domaine de l'analyse de capteurs et de flots de données où les comportements fréquents servent à la surveillance ou à la détection / prévention de pannes.

Nous nous intéressons à la découverte de connaissances au moyen de *règles graduelles*. Les règles graduelles modélisent des co-variations fréquentes sur les valeurs d'items, et sont de la forme "*plus (moins)* A_1 *et ... plus (moins)* A_n*, alors plus (moins)* B_1 *et ... plus (moins)* B_p". Ces règles suscitent depuis quelques années un intérêt croissant (Hüllermeier (2002); Berzal et al. (2007)). La notion de gradualité, et plus particulièrement de règles graduelles, a majoritairement été étudiée dans le domaine du flou. Celles-ci étaient utilisées dans le but de modéliser des systèmes experts. L'accent n'est alors pas mis sur la manière de les extraire, mais

plutôt sur leur rôle au sein de systèmes inductifs (par exemple, “plus le mur est proche, plus le train doit actionner le frein”). Les dépendances graduelles sont mesurées à l'aide d'implication floue, dont plus utilisée est Resher-Gaines, qui assure que le degré d'implication de X est contraint par le degré d'implication de Y.

Bosc et al. (1999) calculent toutes les règles graduelles dans le cadre de la génération de résumés. Les auteurs ne sont pas intéressés par l'extraction de telles règles, mais plutôt par les connaissances qu'elles apportent. Ainsi, différents tests de cohérence sont proposés, permettant par exemple de révéler des zones vides de la base de données.

Hüllermeier (2002) remplace les *tables de contingence* représentant des règles d'associations par des *diagrammes de contingence*. Puis, les corrélations entre variations sont extraites à l'aide d'une régression linéaire directement appliquée sur les diagrammes. Les coefficients de pente et de qualité de la régression sont utilisés afin de décider de la validité d'une règle.

Afin d'éviter des jointures trop coûteuses, Berzal et al. (2007) proposent l'utilisation de l'algorithme Apriori. Ainsi, les itemsets graduels sont définis à l'aide des opérateurs $\{<, >\}$, et la base de données est transformée en une base de couples d'objets. La fouille s'effectue alors directement à partir des couples. Cette méthode est la première permettant de prendre en compte des conjonctions de variations, aussi bien croissantes que décroissantes, dans la condition et la conclusion de la règle.

Enfin, Di Jorio et al. (2008) proposent une première approche efficace en temps, mais non exhaustive car basée sur une heuristique : le nombre d'items dans les règles est augmenté étape par étape en supprimant à chaque passe les plus grands ensembles d'objets de la base qui empêchent la gradualité.

2 Un formalisme pour les itemsets graduels

2.1 Définitions

Personne	**Age (A)**	**Salaire (S)**	**Crédit (C)**
p_1	22	1200	4
p_2	28	1850	5
p_3	24	1200	3
p_4	35	2200	2

TAB. 1: Base exemple $\mathcal{BD}$

Les règles d'association graduelles décrivent des corrélations fréquentes entre variations. Contrairement aux approches par comptage classiques, les variations ne se mesurent pas à partir d'un seul objet, mais à partir d'un ensemble d'objets. Deux types de variations peuvent être considérées : soit la valeur d'un attribut augmente d'un objet à l'autre, soit elle diminue. Ainsi, un item doit être vérifié deux fois, c'est pourquoi nous définissons les items graduels de la manière suivante :

Définition 1. *Soit $\mathcal{I}$ un ensemble d'items, $i \in \mathcal{I}$ un item et $* \in \{\geq, \leq\}$ un opérateur de comparaison. Un item graduel i^* est défini comme un item i associé à un opérateur $*$.*

Par exemple, à partir du tableau 1, six items graduels peuvent être considérés : $\{A^{\geq}, A^{\leq}, S^{\geq}, S^{\leq}, C^{\geq}, C^{\leq}\}$. Le premier item, l'âge, nous amène à considérer deux items graduels : $\{A^{\geq}, A^{\leq}\}$ signifiant respectivement "*l'âge augmente*" et "*l'âge diminue*". Un itemset graduel est alors défini comme un ensemble d'items graduels comme par exemple $S_1 = (A^{\geq}S^{\geq}C^{\leq})$. Celui-ci est obtenu par comparaison entre les propriétés de chaque objet : nous avons comparé les variations entre les attributs d'un objet à l'autre. Habituellement, l'intérêt d'une règle est mesuré par son support, qui reflète la proportion d'objets de la base contenant cette règle. Cette notion est différente dans le cas de la gradualité, car il ne s'agit plus de comptabiliser un nombre d'objets supportant l'itemset, mais le nombre d'objets respectant la variation décrite par l'itemset. A partir de la base de la table 1, nous pouvons donner deux ensembles d'objets respectant S_1 : $\{p_1p_3p_4\}$ et $\{p_2p_4\}$. La fréquence d'un itemset graduel est calculée à partir de l'ensemble le plus représentatif, c'est-à-dire l'ensemble ayant le plus d'éléments :

Définition 2. *Soit $s = (i_1^{*_1}...i_n^{*_n})$ un itemset graduel et G_s l'ensemble des objets respectant s. La fréquence (ou support) de s est donnée par $Freq(s) = \frac{max(|G_s^i|)}{|\mathcal{O}|}$ où $G_s^i \subseteq G_s$ et $\mathcal{O}$ est l'ensemble des objets décrivant la base de données.*

Nous obtenons $Freq(S_1) = \frac{3}{4} = 0.75$, ce qui signifie que S_1 est supporté par 75% de la base. Cependant, la recherche de l'ensemble le plus représentatif revient à calculer tous les ordres possibles de la base, d'autant plus qu'il peut y avoir plusieurs ensembles représentatifs. Ce problème est lié à la fouille d'ordre, dont la solution la plus efficace passe par l'utilisation d'un *ordre total*. Dans la section suivante, nous expliquons comment utiliser les treillis pour extraire l'ensemble le plus représentatif.

2.2 Des treillis pour la gradualité

Les treillis sont utilisés pour représenter des relations d'ordre. Une relation d'ordre partiel sur un ensemble E est une relation binaire réflexive, transitive, et antisymétrique. De plus, si pour tout $(x, y) \in E^2$ nous avons $x \leq y$ ou $y \leq x$, alors E est totalement ordonné. Habituellement, les éléments de E munis d'une relation d'ordre sont organisés sous la forme d'un treillis. Les objets d'une base de données $\mathcal{BD}$ représentant des itemsets graduels peuvent être représentés à l'aide de treillis. De manière plus générale, une relation d'ordre total est définie pour chaque itemset comme suit :

Définition 3. *Soit o et o' deux objets d'une base de données $\mathcal{BD}$ définie sur un ensemble d'items $\mathcal{I} = \{i_1...i_n\}$. $\mathcal{R}_g$ est une relation d'ordre total sur $\mathcal{BD}$ si $\forall j \in [1, n](o[i_j] \geq o'[i_j])$ ou $(o[i_j] \leq o'[i_j])$. On note $o\mathcal{R}_g o'$.*

La recherche de l'ensemble le plus représentatif s'exprime alors différemment en utilisant les treillis. En effet, celui-ci est directement lié à la recherche de *chaîne* :

Définition 4. *Des éléments $a_1, a_2, ..., a_n \in E$ tels que $a_i \mathcal{R} a_{i+1}, \forall i \in [1...n-1]$ forment un ensemble appelé chaîne.*

Le treillis associé à l'itemset graduel S_1 est constitué des deux chaînes représentant l'ensemble des solutions : $\{(p_1p_3p_4), (p_2p_4)\}$. Nous appelons la chaîne composée du plus grand nombre d'éléments *chaîne maximale*. La notion de fréquence est alors directement liée à la cardinalice des chaînes maximales.

Définition 5. *Soit* $s = (i_1^{*_1}...i_n^{*_n})$ *un itemset graduel, et* $\mathcal{L}_s$ *le treillis associé. Soit* $\mathcal{C} = \{C_1, ..., C_p\}$ *l'ensemble des chaînes composant* $\mathcal{L}_s$*. Alors* $Freq(s) = \frac{max(|C_i|)}{|\mathcal{O}|}$*, où* $C_i \in [1, p]$.

Un itemset graduel est fréquent si son support dépasse un seuil minimal défini par l'utilisateur. L'extraction d'itemsets graduels consiste donc en la recherche de l'ensemble des itemsets graduels fréquents à partir d'une base de données contenant des attributs numériques. Afin de prévenir le problème de l'explosion combinatoire, nous utilisons les itemsets graduels complémentaires afin de diminuer l'espace de recherche.

2.3 Espace de recherche graduel

Nous proposons d'adopter des méthodes d'extraction classiques à la problématique d'extraction graduelle. L'une des méthodes les plus adaptées est la méthode "générer-élaguer", dont l'algorithme le plus connu est Apriori (Agrawal et Srikant (1994)).

Une propriété importante des règles graduelles est leur complémentarité. Cette notion, initialement décrite dans Berzal et al. (2007) est formalisé par la définition suivante :

Définition 6. *(itemset graduel complémentaire) Soit* $s = (i_1^{*_1}...i_n^{*_n})$ *un itemset graduel. Son itemset graduel complémentaire est* $c(s) = (i_1^{'*_1}...i_n^{'*_n})$ *si* $\forall j \in [1, n] i_j = i_j^{'}$ *et* $*_j = c_*(*_j^{'})$*, où* $c_*(\geq) = \leq$ *et* $c_*(\leq) = \geq$.

Proposition 1. *i) Soit* s *et* s' *deux itemsets graduels tels que* $c(s) = s'$*. Alors l'ensemble des chaînes composant* $\mathcal{L}_s$ *sont les mêmes que celles composant* $\mathcal{L}_{s'}$*. ii)* $Freq(s) = Freq(c(s))$

Grâce à la proposition 1 ii), la moitié seulement des itemsets seront générés, puisque les autres peuvent être déduit automatiquement. Cette optimisation réduit suffisamment l'espace de recherche afin de passer à l'échelle. Il est maintenant nécessaire de redéfinir l'opération de jointure pour les itemsets graduels. Dans notre contexte, cette opération revient à définir la jointure entre deux treillis.

3 L'algorithme GRITE

3.1 La jointure

Les treillis organisent les objets respectant les variations décrites par un itemset graduel. Ainsi, la jointure de deux treillis $\mathcal{L}_{S_i}$ et $\mathcal{L}_{S_{i+1}}$ doit conserver d'une part les objets communs entre chaque treillis, et d'autre part les ordres communs (conjonctions de variations). Cela revient à calculer toutes les sous-séquences communes aux deux treillis.

L'opération de jointure sera effectuée un nombre exponentiel de fois. Il est donc indispensable d'utiliser une structure de modélisation adaptée. C'est pourquoi nous utilisons les représentations binaires de treillis. En effet, ce type de représentation présente le double avantage d'être peu consommateur de mémoire (un octet représente huit objets) et d'utiliser des opérations binaires très performantes en terme de temps d'exécution. Ainsi, un treillis contenant n sommets peut être projeté en mémoire à l'aide d'une matrice binaire de taille $n \times n$. S'il existe une relation entre un sommet v et un sommet v', alors le bit correspondant à la ligne v et

à la colonne v' vaut 1, et 0 sinon. A partir des matrices binaires, les sous-séquences communes sont celles dont les bits sont à 1 pour chacune, obtenus par l'opération binaire ET entre chaque élément de la matrice :

Théorème 1. *Soit $M_{\mathcal{L}_s}$ la matrice associée au treillis $\mathcal{L}_s$ (resp. $\mathcal{L}_{s'}$) et $M_{\mathcal{L}_{ss'}}$ la matrice associée au treillis $\mathcal{L}_{ss'}$. Nous avons alors la relation suivante : $M_{\mathcal{L}_{ss'}} = M_{\mathcal{L}_s}$ **ET** $M_{\mathcal{L}_{s'}}$.*

Le théorème 1 rend possible l'utilisation des méthodes générer-élaguer dans un temps efficace. En effet, les opérations binaires sont, d'un point de vue processeur, parmi les plus performantes. De plus, la gradualité se mesurant d'un objet à l'autre, les sommets isolés (sommets n'ayant aucune relations) sont élagués permettant ainsi de gagner d'une part de l'espace mémoire, et d'autre part du temps lors de jointures ultérieures.

3.2 Calcul de fréquence

La fréquence d'un itemset graduel s est la longueur de l'une des chaînes maximales du treillis $\mathcal{L}_{ss'}$ associé. Le calcul du plus long chemin est un problème qui peut s'avérer difficile selon les contraintes considérées. Dans notre cas, nous avons un treillis qui peut se ramener à un graphe orienté et acyclique. Les algorithmes de recherche de plus court chemin peuvent alors être appliqués. Cependant, cette classe d'algorithme est polynomiale, et non linéaire. Or, tout comme la jointure, le calcul du support est effectué un nombre exponentiel de fois. C'est pourquoi nous posons la contrainte suivante : chaque sommet ne devra être considéré qu'une seule fois. Pour ce faire, nous avons mis en place un système de "*mémoire*", qui conserve les données obtenues à partir des nœuds de niveau supérieur. Lorsque plusieurs solutions sont possibles, nous conservons le niveau le plus élevé.

La stratégie adoptée est réalisé par un algorithme glouton prennant en entrée un noeud et remplissant la mémoire. Cette mémoire, représentée par un tableau contenant autant d'éléments que de sommets, est préalablement initialisée à -1. Ainsi, lorsqu'un -1 est rencontré, cela signifie que le sommet correspondant n'a pas été visité. Pour chaque fils du nœud en question ayant une mémoire à -1, l'algorithme est récursivement appelé. Lorsqu'une feuille est rencontrée, sa mémoire prend pour valeur 1. Enfin, quand chaque fils a sa mémoire renseignée, le père prend le niveau maximal.

4 Expérimentations

Nous avons, dans un premier temps, testé GRITE sur des jeux de données synthétiques contenant notamment un très large nombre d'attributs comparé au nombre d'objets. Les résultats, en terme de temps et d'occupation mémoire sont encourageants : l'extraction est possible sur des bases dont le format était jusqu'à présent difficile à traiter. Puis, dans un second temps, GRITE a été testé sur un jeu de données réelles concernant la maladie d'Alzheimer. La base contient des notes de chercheurs en psychologie concernant les sentiments et la mémoire. De plus, chaque patient a passé les tests classiques du manuel des démences ("Diagnostic manual of mental disorders"). Les psychologues ont demandé à leurs patients de se souvenir d'un bon moment de leur vie, puis d'un mauvais moment, et enfin d'un moment neutre. Ensuite, ils ont évalué leurs souvenirs concernant les bruits, la disposition spatiales des objets ainsi que leurs sentiments. Ce jeux de données contient 33 patient et 122 attributs. Voici quelques règles extraites et validées comme intéressantes par les experts :

– “Plus le souvenir de la disposition spaciale des personnes et du moment de la journée est bon, alors plus le souvenir de l’endroit est bon” (87.88%, 100%)
– “Plus le nombre d’identifications au test RI 48 augmente et plus le souvenir du moment de la journée et de la disposition spaciale d’un moment neutre est bon, alors plus le test du MMS est réussi” (81.82%, 100%)

La seconde règle s’avère la plus intéressante, car elle lie des données de test à des données “mémoire”. Cela permet de relier l’état de santé mental du patient au type de souvenirs qu’il garde d’un moment neutre.

5 Conclusion

Nous avons proposé une nouvelle approche d’extraction de règles graduelles. Les approches existantes utilisent les sous-ensembles flous qui donnent une description sémantique concrète aux règles extraites, mais ne permettent pas le passage à l’echelle. Notre algorithme, nommé GRITE, tire avantage des treillis et de leur représentation binaire afin de représenter les variations. Les règles extraites à partir d’un jeu de données réelles souligne l’intérêt de telles règles pour les experts. Les résultats obtenus ouvrent de nombreuses perspectives, notamment en terme de rafinement des resultats. Cela passe par la proposition de nouvelles mesures de qualité.

Références

Agrawal, R. et R. Srikant (1994). Fast Algorithms for Mining Association Rules. In *20th International Conference on Very Large Data Bases, (VLDB’94)*, pp. 487–499.

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

Bosc, P., O. Pivert, et L. Ughetto (1999). On data summaries based on gradual rules. In *Proceedings of the 6th International Conference on Computational Intelligence, Theory and Applications*, London, UK, pp. 512–521. Springer-Verlag.

Di Jorio, L., A. Laurent, et M. Teisseire (2008). Fast extraction of gradual association rules : A heuristic based method. In *IEEE/ACM International Conference on Soft computing as Transdisciplinary Science and Technology*, Paris/Cergy Pontoise.

Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. In *PKDD ’02 : Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK, pp. 200–211. Springer-Verlag.

Srikant, R. et R. Agrawal (1996). Mining Quantitative Association Rules in Large Relational Tables. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pp. 1–12.

Summary

Gradual rules have been more and more studied these last years. Such rules can be applied in many domains, as bioinformatic, fuzzy controlers, sensor readings or data streams. In this paper, we tackle the particular problem of handling huge volumes by proposing scalable lattice-based methods.