Extraction de motifs graduels emergents

Anne Laurent, Marie-Jeanne Lesot, Maria Rifqi

Extraction de motifs graduels émergents

Contrasting data using gradual patterns

Anne Laurent[1] Marie-Jeanne Lesot[2] Maria Rifqi[3]
[1] LIRMM - Université Montpellier 2, Montpellier, France
[2] Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
[3] LEMMA - Université Panthéon-Assas, Paris, France
anne.laurent@lirmm.fr, marie-jeanne.lesot@lip6.fr, maria.rifqi@u-paris2.fr

Résumé :
L'extraction de motifs émergents a pour objectif de souligner les caractéristiques distinctives d'une base de données par opposition à une base de référence, afin de mettre en évidence leurs différences. Cet article considère le cas particulier des motifs graduels émergents et vise, donc, à extraire des co-variations d'attributs discriminantes. Il discute les spécificités des motifs graduels nécessitant le développement d'une nouvelle méthode et propose la transposition adaptée d'un algorithme efficace basé sur la notion de bordure, en justifiant son applicabilité au cas des motifs graduels. Il illustre les résultats obtenus sur des données de l'UCI.

Mots-clés :
Motifs graduels, motifs émergents, caractérisation discriminante.

Abstract:
Mining emerging patterns aims at contrasting data sets and identifying itemsets that characterise a data set by contrast to a reference data set, so as to capture and highlight their differences. This paper considers the case of emerging gradual patterns, to extract discriminant attribute co-variations. It discusses the specific features of these gradual patterns and proposes to transpose an efficient border-based algorithm, justifying its applicability to the gradual case. Illustrative results obtained from a UCI data set are described.

Keywords:
Gradual patterns, emerging patterns, discriminant characterisation

1 Introduction

Les motifs graduels [2, 6, 13] extraient de bases de données des connaissances sous la forme de co-variations d'attributs, linguistiquement exprimées comme « plus A est élevé, plus B est élevé », où A et B sont des attributs numériques ou des degrés d'appartenance à des modalités floues. Cet article considère la tâche d'extraction de motifs graduels *émergents*, définis comme des motifs graduels qui permettent de caractériser un ensemble de données *par opposition* à des données de référence, c'est-à-dire qui sont observés dans une base mais non dans l'autre. De tels motifs visent à mettre en évidence les spécificités des données considérées en termes de co-variations d'attributs et à souligner leurs différences par rapport aux données de référence.

A titre d'exemple, on peut considérer des données qui décrivent des utilisateurs de réseaux sociaux par leur âge et les outils qu'ils privilégient : si l'on compare des données obtenues avant et après 2014 respectivement, les motifs graduels émergents pourraient indiquer que le motif « plus les utilisateurs sont jeunes, plus ils utilisent Facebook » est valide avant 2014, puis est remplacé par « plus les utilisateurs sont jeunes, plus ils utilisent Snapchat ». Dans ce cas, les motifs graduels émergents permettent de s'adapter à l'évolution des données au cours du temps et de souligner leurs changements en termes de co-variations d'attributs.

Cette tâche de fouille de données étend aux motifs graduels les notions de motifs émergents [7] et de *contrast sets* [1] : ces derniers visent à identifier des motifs discriminants dans des bases de données transactionnelles. Leur extraction soulève des problèmes de complexité calculatoire, principalement dus à l'absence de propriété d'anti-monotonie de la notion d'émergence : les sous-motifs d'un motif émergent ne sont pas nécessairement émergents. Aussi, ils ne peuvent être identifiés en mettant en œuvre une démarche classique

de type “générer et filtrer” et requièrent des méthodes spécifiques.

Cet article considère la tâche d'extraction de motifs graduels émergents : il propose d'adapter un algorithme efficace proposé pour les motifs classiques qui exploite une représentation compacte des motifs maximaux fréquents [7, 8]. Il justifie sa pertinence pour les motifs graduels et propose une transposition adaptée qui tient compte de leurs spécificités.

La section 2 rappelle les définitions de motifs graduels et de motifs émergents, ainsi que l'algorithme d'extraction MBD-LL BORDER [7]. La section 3 justifie sa transposition au cas des motifs graduels et présente les adaptations proposées. La section 4 présente les résultats expérimentaux obtenus sur la base de données *vehicles* de UCI [17].

2 Motifs graduels et motifs émergents

Cette section présente deux types de motifs particuliers, qui permettent d'extraire différentes formes de connaissances : les motifs graduels et les motifs émergents. Elle rappelle, pour chacun, sa définition, ses spécificités et les algorithmes d'extraction existants.

2.1 Motifs graduels

Alors que les motifs classiques s'appliquent à des données transactionnelles, décrites par des attributs binaires indiquant la présence ou l'absence de chaque item, les motifs graduels [2, 6, 13] sont extraits de données décrites par des valeurs réelles, associées à des attributs numériques ou à des modalités floues. Les motifs graduels sont linguistiquement exprimés sous la forme « plus A est élevé, plus B est élevé ». Ils imposent des contraintes d'ordre sur les valeurs d'attributs à l'ensemble des données, considérées globalement, et ils doivent être distingués des règles graduelles floues [3, 9, 11]. En effet, ces dernières considèrent chaque donnée individuellement, généralisant au cas flou les règles d'association classiques, alors que les motifs graduels expriment une tendance globale à travers l'ensemble des données.

Ci-dessous, $\mathcal{D}$ désigne l'ensemble de données considéré, n le nombre d'objets qu'il contient et m le nombre d'attributs, numériques.

Définitions. Formellement, un item graduel, noté $A^{\geq}$ ou $A^{\leq}$, est défini [2, 6] comme un couple constitué d'un attribut A et d'un sens de variation, $\geq$ ou $\leq$. Il est linguistiquement exprimé comme *plus A est élevé* pour $A^{\geq}$ et *plus A est faible* pour $A^{\leq}$.

Un *motif graduel* M est un ensemble d'items graduels, noté $M = \{A_i^{*_i}, i = 1..k\}$, où $*_i \in \{\geq, \leq\}$ pour tout $i \in [1, k]$. On lui associe sa taille, k, définie comme le nombre d'attributs qu'il implique et le pré-ordre $\preceq_M$ défini sur $\mathcal{D}^2$ comme $x \preceq_M x'$ ssi $\forall A_i^{*_i} \in M, A_i(x) *_i A_i(x')$ où $A_i(x)$ représente la valeur de l'attribut A_i pour l'objet x. M est sémantiquement interprété comme la conjonction de ses items graduels : $A^{\geq}B^{\leq}$ est ainsi linguistiquement exprimé comme *plus A est élevé, plus B est faible*.

Critère de qualité. La qualité d'un motif, définie comme son *support*, indique à quel point il est valide pour un ensemble de données. Deux approches principales peuvent être distinguées pour les motifs graduels. La première tient compte des valeurs d'attributs et repose par exemple sur une régression linéaire [12] : le support d'un motif graduel est mesuré comme la qualité de la régression, combinée à la pente de la droite. Cette approche nécessite de combiner les valeurs numériques des attributs ; elle s'applique en particulier à des données floues pour lesquelles les combinaisons peuvent être effectuées par des $\top$-normes par exemple.

La seconde interprétation considère l'ordre induit par les attributs, en ignorant leurs valeurs. L'approche par ensembles d'objets compatibles [5, 6] identifie des sous-ensembles d'objets qui peuvent être ordonnés de telle sorte que tous les couples d'objets satisfont l'ordre

induit par le motif. Formellement,

$$supp(M) = \frac{1}{|\mathcal{D}|} \max_{\mathcal{D}^* \in \mathcal{L}(M)} |\mathcal{D}^*| \quad (1)$$

où $\mathcal{L}(M)$ est l'ensemble de tous les sous-ensembles maximaux $\mathcal{D}^* = \{x_1, ... x_m\} \subseteq \mathcal{D}$ pour lesquels il existe une permutation π telle que $\forall l \in [1, m-1]$, $x_{\pi_l} \preceq_M x_{\pi_{l+1}}$.

L'approche par corrélation d'ordres [2, 13] considère une vue plus locale, basée sur les couples de données et non des sous-ensembles. Elle est basée sur un comptage des couples qui satisfont l'ordre induit par le motif :

$$supp(M) = \frac{|\{(x, x') \in \mathcal{D}^2 / x \preceq_M x'\}|}{|\mathcal{D}|(|\mathcal{D}|-1)/2} \quad (2)$$

Malgré les différences d'interprétation [4], ces définitions de support satisfont la propriété classique d'anti-monotonie, permettant la définition d'algorithmes efficaces d'extraction de motifs graduels fréquents.

Spécificités. Deux particularités des motifs graduels par rapport aux motifs classiques doivent être soulignées : à la fois en termes de données et d'attributs, ils s'appliquent à des paires et non individuellement.

En effet, comme l'indique l'ordre qu'ils induisent, les motifs graduels s'appliquent aux paires de données, ce qui augmente significativement la complexité de calcul de leur traitement : ils peuvent être interprétés comme des motifs classiques issus d'une base de données modifiée, qui transforme les données initiales en construisant une transaction pour chaque couple de données [2]. La mise en œuvre explicite de cette transformation [2] nécessite des approximations pour réduire les coûts de calcul en pratique. Une autre approche consiste à représenter les données par des matrices de concordance, qui indiquent, pour chaque couple de données, s'il satisfait un motif graduel donné. Ces matrices, à valeurs booléennes, permettent un traitement efficace par des opérations binaires [6, 13].

Outre cette spécificité portant sur les paires de données, les motifs graduels s'appliquent à des paires d'attributs : un motif graduel élémentaire est de taille 2. En effet, un item graduel de taille 1 n'impose pas de contrainte, puisque toute paire de données peut être ordonnée trivialement pour respecter l'ordre qu'il induit. Aussi, l'approche par transformation explicite des données [2] construit des items pour toutes les paires d'items graduels. Elle conduit donc à une base transformée contenant $n(n-1)/2$ lignes, une pour chaque paire de données, et $m(m-1)$ colonnes, pour représenter tous les 2-items graduels [1] $A^{\geq}B^{\geq}$ et $A^{\geq}B^{\leq}$ pour chaque paire d'attributs A, B.

Ces spécificités imposent des contraintes sur l'adaptation des algorithmes d'extraction de motifs émergents au cas des motifs graduels, comme discuté dans la section 3.

2.2 Motifs émergents

Dans le cadre de la fouille de données transactionnelles, les motifs émergents [7] sont définis comme des motifs qui caractérisent un ensemble de données *par opposition* à des données de référence. Leur extraction permet d'identifier des caractéristiques discriminantes en mettant en évidence des motifs valables pour une base et non l'autre.

Définition. Formellement, en notant $\mathcal{D}_1$ et $\mathcal{D}_2$ deux bases de données et ρ une valeur de seuil, un motif M est *émergent* dans $\mathcal{D}_2$ par opposition à $\mathcal{D}_1$ si son support augmente significativement [7] :

$$\frac{supp_{\mathcal{D}_2}(M)}{supp_{\mathcal{D}_1}(M)} \geq \rho \quad (3)$$

Ce rapport est appelé *taux de croissance*. Si le support de M vaut 0 sur $\mathcal{D}_1$ et sur $\mathcal{D}_2$, on considère que ce taux est nul.

1. Les motifs graduels $A^{\leq}B^{\leq}$ et $A^{\leq}B^{\geq}$ peuvent être considérés comme équivalents à $A^{\geq}B^{\geq}$ et $A^{\geq}B^{\leq}$ respectivement, car ils induisent l'ordre inverse et sont satisfaits par les mêmes paires de données.

Spécificités. La définition des motifs émergents fait intervenir deux bases de données, ce qui augmente la complexité calculatoire par rapport aux motifs classiques. De plus, le critère de taux de croissance précédent ne vérifie pas de propriété d'anti-monotonie : les sous-motifs d'un motif dont le taux de croissance est supérieur au seuil peuvent ne pas vérifier cette propriété. Aussi, l'utilisation de méthodes reposant sur des extensions des algorithmes classiques d'extraction, de type APRIORI, est exclue. Il faut souligner que cette absence de monotonie vient de façon générale de la définition même de motifs émergents, au-delà de celle du critère de taux de croissance.

Algorithmes d'extraction. Les motifs émergents peuvent être interprétés comme des motifs qui sont rares dans $\mathcal{D}_1$ mais fréquents dans $\mathcal{D}_2$. Une approche consistant à éliminer des candidats établis à partir de $\mathcal{D}_2$ selon leur rareté dans $\mathcal{D}_1$ ne peut néanmoins pas être mise en œuvre, en raison du très grand nombre de motifs rares et de l'absence de monotonie.

Une interprétation proche consiste à considérer que l'on extrait les motifs fréquents des deux bases $\mathcal{D}_1$ et $\mathcal{D}_2$ puis que l'on conserve uniquement ceux qui sont présents dans les résultats de $\mathcal{D}_2$ et absents de ceux de $\mathcal{D}_1$, c'est-à-dire que l'on calcule la différence ensembliste de ces résultats. Toutefois, en raison du nombre élevé de motifs fréquents, cette approche nécessite à la fois une représentation compacte des motifs fréquents et un mode de calcul efficace de leurs différences ensemblistes. Il a ainsi été proposé d'utiliser les motifs maximaux [7, 8] ou les motifs fermés [18, 19]. L'exploitation des premiers, sur laquelle repose la méthode que nous proposons, est détaillée dans la section suivante.

Les méthodes proposées dans le domaine des *contrast sets* [1, 10, 15] suivent un principe différent : elles expriment la tâche comme l'identification de différences significatives entre les distributions des motifs dans les deux bases et exploitent des tests statistiques pour identifier les motifs discriminants.

On peut noter que les notions de motifs émergents et de *contrast sets* peuvent être rapprochées des motifs graduels caractérisés [16] : ceux-ci extraient automatiquement des sous-ensembles de données sur lesquels le support d'un motif graduel est supérieur au support obtenu sur l'intégralité des données. Les motifs émergents considèrent un processus inverse d'identification de motifs pour des ensembles de données fixés.

2.3 Extraction par bordure de motifs émergents

Cette section détaille l'approche d'extraction des motifs émergents qui exploite une représentation compacte des motifs fréquents basée sur les motifs maximaux, appelée bordure [7, 8].

Définition. Une bordure est définie [7, 8] comme un couple $\langle \mathcal{L}, \mathcal{R} \rangle$ d'antichaînes [2] tel que chaque élément de $\mathcal{L}$ est un sous-ensemble d'un élément de $\mathcal{R}$, et chaque élément de $\mathcal{R}$ est un sur-ensemble d'un élément de $\mathcal{L}$. On dit qu'une bordure est *ancrée à gauche* lorsque $\mathcal{L} = \{\emptyset\}$.

La bordure $\langle \mathcal{L}, \mathcal{R} \rangle$ représente de manière compacte la collection $\{Y | \exists X \in \mathcal{L}, \exists Z \in \mathcal{R}$ tels que $X \subseteq Y \subseteq Z\}$: c'est une représentation concise qui évite d'énumérer explicitement tous ces ensembles.

Les bordures sont particulièrement utiles dans le cas de collections convexes : la propriété-clé d'une collection convexe d'ensembles $\mathcal{S}$ est qu'elle peut être décrite de manière unique par une bordure. Plus précisément, sa bordure $\langle \mathcal{L}, \mathcal{R} \rangle$ est telle que $\mathcal{L}$ (resp. $\mathcal{R}$) est la collection des ensembles minimaux (resp. maximaux) dans $\mathcal{S}$. On peut noter qu'une collection non convexe peut être décomposée en une union de collections convexes [8].

Exploitation des bordures. La pertinence de la représentation compacte par bordure vient de

2. Une collection d'ensembles $\mathcal{S}$ est une antichaîne si $\forall X, Y \in \mathcal{S}$, $X \not\subseteq Y$ et $Y \not\subseteq X$.

ce que de nombreuses opérations ensemblistes peuvent être exprimées en termes de bordure, sans nécessiter l'énumération explicite de tous les ensembles contenus dans les collections considérées. En particulier, l'algorithme BORDER-DIFF [7], optimisé dans [8], permet de calculer la différence ensembliste de deux collections représentées par des bordures ancrées à gauche en exploitant uniquement cette représentation.

Application à l'extraction de motifs émergents. On peut montrer [7] qu'un ensemble de motifs fréquents peut être représenté sous le forme d'une bordure ancrée à gauche, plus précisément sous la forme $\langle\{\emptyset\}, \mathcal{R}\rangle$ où $\mathcal{R}$ est l'ensemble des motifs fréquents maximaux. Aussi, la différence ensembliste entre deux ensembles de motifs fréquents peut être calculée de manière efficace.

L'algorithme MBD-LL-BORDER [7] met en œuvre cette approche : une première étape consiste à extraire les motifs fréquents de $\mathcal{D}_1$ et $\mathcal{D}_2$ en utilisant un algorithme classique et à les représenter par leurs bordures. La seconde applique BORDER-DIFF pour calculer leur différence ensembliste, fournissant une représentation par bordure de tous les motifs émergents.

3 Extraction de motifs graduels émergents

Cette section décrit l'approche que nous proposons pour extraire des motifs graduels émergents, basée sur l'algorithme MBD-LL-BORDER [7] : après avoir discuté des propriétés des motifs graduels qui justifient ce choix, la définition de bordure graduelle ainsi que la transposition de l'algorithme au cas des motifs graduels sont décrites.

3.1 Motivation de l'approche par bordure

Dans le but d'adapter les algorithmes d'extraction de motifs émergents, comme ceux mentionnés dans la section 2.2, au cas graduel, il est nécessaire d'examiner si les motifs graduels satisfont les conditions respectives des algorithmes, en tenant compte de leurs spécificités rappelées dans la section 2.1.

Tout d'abord, on peut remarquer que la transposition de la notion de motif fermé au cas graduel est délicate : la fermeture d'un motif classique est définie comme l'intersection des transactions qui le contiennent. Cette définition peut trivialement être adaptée au cas graduel si l'ensemble des données est transformé dans une forme transactionnelle, comme proposé dans [2] ; cependant, cette approche a un coût calculatoire élevé. On peut envisager une approche basée sur l'interprétation par sous-ensembles de données compatibles [6] : elle identifie les données qui satisfont le motif graduel, de la même manière que la définition de la fermeture identifie d'abord les sous-ensembles de transactions qui contiennent le motif classique. Toutefois, la définition et le calcul de leur intersection, pour produire l'ensemble de tous les motifs graduels que les données ont en commun, sont plus complexes que l'interprétation ensembliste des motifs classiques.

De même, l'approche par *contrast set*, basée sur la notion de distribution de motifs, peut être directement transposée aux données réécrites sous forme transactionnelle. Cependant, elle soulève le même problème de coût de calculs.

Dans les deux cas, la difficulté vient du fait que les motifs graduels s'appliquent à des paires de données, conduisant à une complexité quadratique. Au contraire, l'approche par bordure apparaît comme plus réalisable du fait des propriétés des motifs graduels. En effet, en raison de la propriété d'anti-monotonie satisfaite par les 3 définitions de support rappelées dans la section 2.1, on peut montrer que

Propriété 1 *L'ensemble des motifs graduels fréquents est convexe.*

Aussi, dans les 3 cas, l'ensemble des motifs

graduels fréquents peut être représenté par des bordures et l'algorithme MBD-LL-BORDER [7] peut être appliqué pour extraire les motifs émergents graduels.

3.2 Représentation par bordure des motifs graduels

Du fait de la propriété-clé précédente, les motifs graduels peuvent être à la fois représentés de manière compacte et traités à travers la bordure associée à leurs motifs maximaux.

Définition de bordure graduelle. Pour extraire et faciliter la manipulation des bordures associées aux motifs graduels maximaux, nous proposons d'utiliser la transformation explicite suivante : un motif graduel $M = \{A_i^{*i}, i = 1..k\}$ est représenté comme l'ensemble de tous les motifs graduels de taille 2 qu'il contient, $\{(A_{i_1}^{*i_1}, A_{i_2}^{*i_2}), (i_1, i_2) \in \{1..k\}^2, i_1 < i_2\}$. Par exemple, le 4-motif graduel $A^{\geq}B^{\geq}C^{\leq}D^{\geq}$ est représenté comme l'ensemble des 6 motifs graduels $\{A^{\geq}B^{\geq}$, $A^{\geq}C^{\leq}$, $A^{\geq}D^{\geq}$, $B^{\geq}C^{\leq}$, $B^{\geq}D^{\geq}$, $C^{\leq}D^{\geq}\}$. Dans le cas classique, le 4-motif $ABCD$ est décomposé en ses 4 items A, B, C et D, ce qui illustre à nouveau la complexité accrue des motifs graduels.

Plus généralement, un motif graduel de taille k est donc représenté par l'ensemble de $k(k-1)/2$ motifs graduels de taille 2, chacun d'entre eux étant codé avec un unique identifiant. Cette étape a un coût non négligeable, mais inférieur à celui de la réécriture des données en une base transactionnelle. De plus, elle simplifie l'application de l'algorithme MBD-LL-BORDER.

Aussi, nous proposons la représentation par bordure des motifs graduels suivante :

Propriété 2 *Une collection $\mathcal{S}$ de motifs graduels fréquents est représentée comme une bordure ancrée à gauche $\langle\{\emptyset\}, \mathcal{R}\rangle$, où $\mathcal{R}$ est l'ensemble des itemsets graduels maximaux dans $\mathcal{S}$, où chacun d'eux est représenté par l'ensemble de ses motifs graduels de taille 2.*

Exploitation des bordures graduelles. L'exploitation de cette représentation par bordure graduelle présente une spécificité à souligner, quant au calcul de l'union ensembliste. En effet, l'union de 2 motifs graduels n'est pas l'union ensembliste de leurs composants de taille 2, car ces derniers ne sont pas indépendants : certains sont implicites et doivent explicitement ajoutés. Ainsi, dans le cas élémentaire de motifs graduels de taille 2, l'union de $A^{\geq}B^{\leq}$ avec $B^{\leq}C^{\geq}$ est le motif graduel de taille 3 $A^{\geq}B^{\leq}C^{\geq}$, correspondant à 3 motifs graduels de taille 2 : elle ne contient pas seulement l'union des 2 motifs considérés mais aussi, implicitement, $A^{\geq}C^{\geq}$.

3.3 Algorithme proposé

L'adaptation de l'algorithme MBD-LL-BORDER [7] au cas des motifs graduels prend donc la forme simple suivante :

1. Extraire les motifs graduels fréquents de $\mathcal{D}_1$ et $\mathcal{D}_2$, avec des seuils de support respectifs s_1 and s_2, en utilisant l'une des définitions du support rappelées en section 2.1.

2. Etablir leur représentation par bordure, comme décrit dans la sous-section précédente.

3. Appliquer MBD-LL-BORDER à ces bordures, en calculant l'union comme indiqué dans la sous-section précédente.

Cette méthode produit une représentation par bordure des motifs graduels émergents.

4 Résultats expérimentaux

Afin d'illustrer la méthode proposée, cette section présente les résultats obtenus sur une partie de la base de données *vehicles* de UCI [17]. Celle-ci décrit des vues de 2 types de véhicule, des camionnettes et des bus en utilisant 18 attributs de forme, dont les noms sont donnés dans le tableau 1. Le but est d'extraire les contraintes de co-variations d'attributs qui s'appliquent à un type de véhicules mais non à l'autre.

Lors de la première étape de la méthode, pour

1	compactness
2	circularity
3	distance circularity
4	radius ratio
5	pr.axis aspect ratio
6	max.length aspect ratio
7	scatter ratio
8	elongatedness
9	pr.axis rectangularity
10	max.length rectangularity
11	scaled variance along major axis
12	scaled variance along minor axis
13	scaled radius of gyration
14	skewness about major axis
15	skewness about minor axis
16	kurtosis about minor axis
17	kurtosis about major axis
18	hollows ratio

Tableau 1 – Attributs de forme de la base UCI *vehicles* (voir [17] pour leurs définitions)

extraire les itemsets graduels fréquents, nous utilisons l'algorithme GRAANK [13] qui met en œuvre une interprétation des motifs graduels en termes de corrélation d'ordres, en utilisant les matrices de concordance comme représentation efficace des données.

Nous extrayons les motifs graduels émergents en considérant pour $\mathcal{D}_1$ les 198 vues de camionnettes et pour $\mathcal{D}_2$ les 217 vues de bus. Le tableau 2 présente les bordures obtenues, d'abord ordonnées par la taille de $\mathcal{R}$ puis par les items apparaissant dans $\mathcal{L}$. Pour simplifier les notations, $A^{\geq}$ (resp. $A^{\leq}$) est noté A+ (resp. A-).

Lorsque l'on cherche à caractériser les bus par opposition aux camionnettes avec un seuil de support de 0,75 dans les 2 ensembles, on obtient 3 bordures. Pour chacune d'elles $\mathcal{L} = \mathcal{R}$: chaque bordure contient en fait un seul motif graduel. Tous les motifs sont de taille 2. En particulier, on remarque que les bus satisfont la covariation « plus il est compact, moins il est allongé », contrairement aux camionnettes.

De manière asymétrique, les camionnettes semblent avoir plus de caractéristiques spécifiques par rapport aux bus, puisque leurs motifs graduels émergents fournissent une image plus complexe. En effet, 7 bordures sont obtenues ; de plus pour certaines, on a $\mathcal{L} \neq \mathcal{R}$. Plus précisément, 3 bordures avec $\mathcal{R}$ de taille 2

Bus vs camionnettes, $s_1 = s_2 = 0,75$	
$\mathcal{L}_1$ = {(1+ 8-)}	$\mathcal{R}_1$ = {(1+ 8-)}
$\mathcal{L}_2$ = {(1+ 12+)}	$\mathcal{R}_2$ = {(1+ 12+)}
$\mathcal{L}_3$ = {(4+ 12+)}	$\mathcal{R}_3$ = {(4+ 12+)}
Camionnettes vs bus, $s_1 = s_2 = 0,75$	
$\mathcal{L}_1$ = {(2+ 11+)}	$\mathcal{R}_1$ = {(2+ 11+)}
$\mathcal{L}_2$ = {(4+ 5+)}	$\mathcal{R}_2$ = {(4+ 5+)}
$\mathcal{L}_3$ = {(10+ 11+)}	$\mathcal{R}_3$ = {(10+ 11+)}
$\mathcal{L}_4$ = {(2+ 7+), (2+ 12+)}	$\mathcal{R}_4$ = {(2+ 7+ 12+)}
$\mathcal{L}_5$ = {(6+ 7+), (6+ 12+)}	$\mathcal{R}_5$ = {(6+ 7+ 12+)}
$\mathcal{L}_6$ = {(8- 10+), (7+ 10+), (10+ 12+)}	
	$\mathcal{R}_6$ = {(7+ 8- 10+ 12+)}
$\mathcal{L}_7$ = {(3+ 7+ 8-)}	$\mathcal{R}_7$ = {(3+ 7+ 8- 12+)}
Camionnettes vs bus, $s_1 = 0,5$, $s_2 = 0,75$	
$\mathcal{L}_1$ = {(6+ 7+), (6+ 12+)}	$\mathcal{R}_1$ = {(6+ 7+ 12+)}

Tableau 2 – Representation par bordures des motifs graduels émergents. $A^{\geq}$ (resp. $A^{\leq}$) est noté A+ (resp. A-). La signification des attributs est donnée dans le tableau 1.

sont obtenues également, mais aussi 2 bordures avec $\mathcal{R}$ de taille 3 et 2 bordures de taille 4. Pour la bordure $\langle \mathcal{L}_7, \mathcal{R}_7 \rangle$ par exemple, $\mathcal{R}_7$ indique que le motif graduel (3+ 7+ 8- 12+) a un support supérieur à 0,75 pour les camionnettes mais inférieur pour les bus. De plus, aucun de ses sous-motifs de taille 3 n'est émergent pour les camionnettes, excepté (3+ 7+ 8-), car ils sont tous exclus de $\mathcal{L}_7$: la spécificité des camionnettes vient de la combinaison de ces 3 items. Aucun des sous-motifs de taille 2 n'est spécifique des camionnettes. Le même type de commentaire s'applique aux autres bordures.

Pour accroître le pouvoir de discrimination des motifs graduels émergents, le seuil de support pour $\mathcal{D}_1$, s_1, peut être diminué, de manière à se concentrer sur les motifs qui sont plus rares dans $\mathcal{D}_1$ et par conséquent plus émergents dans $\mathcal{D}_2$. La partie inférieure du tableau 2 montre que, pour $s_1 = 0,5$ et $s_2 = 0,75$, une seule bordure est observée, contenant le motif de taille 3 (6+ 7+ 12+) et tous ses sous-motifs, excepté (7+ 12+). Cette bordure est, naturellement, aussi présente dans les résultats obtenus avec $s_1 = s_2 = 0,75$.

Lorsque l'on oppose les bus aux camionnettes avec les même seuils de support, aucune bordure, et donc aucun motif émergent, n'est obtenue. Ceci est cohérent avec le faible nombre de bordures observées avec $s_1 = s_2 = 0,75$: il semble plus difficile d'opposer les camionnettes aux bus, en termes de co-variation d'attributs, que l'inverse. Ce résultat peut être interprété en termes de compacité et séparabilité, ou de typicalité [14], de ces 2 classes.

L'originalité des motifs graduels émergents par rapport aux autres caractérisations discriminantes de classes, comme les méthodes de classification, vient de la comparaison spécifique considérée entre classes : elle repose sur les ordres induits par les attributs *i.e.* les tendances globales, et non sur des similarités ou des distances entre des exemples des classes, ni sur les distributions des valeurs des attributs.

5 Conclusion et perspectives

Cet article a présenté une approche pour extraire des motifs graduels émergents, rendant possible l'opposition de bases de données en termes de co-variations d'attributs. Elle repose sur la transposition, aux motifs graduels, de la représentation compacte par bordure d'une collection, et sur le calcul efficace qu'elle permet.

Les travaux futurs visent à enrichir l'étude expérimentale de l'approche proposée, en particulier en termes de passage à l'échelle, pour mesurer son applicabilité à de grandes bases de données. La difficulté principale vient de l'interprétabilité plus que de la complexité calculatoire : le problème crucial est de définir une représentation d'une collection de motifs graduels émergents, qui peut être de très grande taille, de telle sorte qu'elle permette à des experts de comprendre la connaissance extraite. Elle pourra s'appuyer sur la conception d'outils de visualisation dédiés.

Références

[1] S. Bay and M. Pazzani. Detecting group differences : mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3) :213–246, 2001.

[2] F. Berzal, J.-C. Cubero, D. Sanchez, M.-A. Vila, and J. M. Serrano. An alternative approach to discover gradual dependencies. *IJUFKS*, 15(5) :559–570, 2007.

[3] B. Bouchon-Meunier and S. Desprès. Acquisition numérique / symbolique de connaissances graduelles. In *3èmes Journées Nationales du PRC Intelligence Artificielle*, pages 127–138, 1990.

[4] B. Bouchon-Meunier, A. Laurent, M.-J. Lesot, and M. Rifqi. Strengthening fuzzy gradual rules through "all the more" clauses. In *Proc. of fuzzIEEE'10*, pages 2940–2946, 2010.

[5] L. Di Jorio, A. Laurent, and M. Teisseire. Fast extraction of gradual association rules : a heuristic based method. In *Proc. of CSTST'08*, 2008.

[6] L. Di Jorio, A. Laurent, and M. Teisseire. Mining frequent gradual itemsets from large databases. In *Proc. of IDA'09*, 2009.

[7] G. Dong and J. Li. Efficient mining of emerging patterns : Discovering trends and differences. In *Proc. of KDD'99*, 1999.

[8] G. Dong and J. Li. Mining border descriptions of emerging patterns from dataset pairs. *Knowledge and Information Systems*, 5 :178–202, 2005.

[9] D. Dubois and H. Prade. Gradual inference rules in approximate reasoning. *Information Sciences*, 61(1-2) :103–122, 1992.

[10] R. Hilderman and T. Peckham. A statistically sound alternative approach to mining contrast sets. In *Proc. of the 4th Australia Data Mining Conf.*, pages 157–172, 2005.

[11] E. Hüllermeier. Implication-based fuzzy association rules. In *Proc. of PKDD'01*, pages 241–252, 2001.

[12] E. Hüllermeier. Association rules for expressing gradual dependencies. In *Proc. of PKDD'02*, pages 200–211. Springer-Verlag, 2002.

[13] A. Laurent, M.-J. Lesot, and M. Rifqi. GRAANK : exploiting rank correlations for extracting gradual itemsets. In *Proc. of FQAS'09*, pages 382–393, 2009.

[14] M.-J. Lesot, M. Rifqi, and B. Bouchon-Meunier. Fuzzy prototypes : From a cognitive view to a machine learning principle. In *Fuzzy Sets and Their Extensions : Representation, Aggregation and Models*, pages 431–452. Springer, 2007.

[15] J. Lin and E. Keogh. Group SAX : extending the notion of contrast sets to time series and multimedia data. In *Proc. of PKDD'06*, pages 284–296, 2006.

[16] A. Oudni, M.-J. Lesot, and M. Rifqi. Characterisation of gradual itemsets through "especially if" clauses based on mathematical morphology tools. In *Proc. of EUSFLAT'13*, pages 826–833, 2013.

[17] J. Siebert. Vehicle recognition using rule based methods. Technical report, Turing Institute, Glasgow, 1987.

[18] A. Soulet, B. Crémilleux, and F. Rioult. Condensed representation of emerging patterns. In *Proc. of PAKDD'04*, pages 127–132, 2004.

[19] A. Soulet, B. Crémilleux, and F. Rioult. Condensed representation of EPs and patterns quantified by frequency based measures. In *Proc. of KDID'04*, pages 173–189, 2005.