Extraction de motifs graduels par corrélations d'ordres induits

Anne Laurent, Marie-Jeanne Lesot, Maria Rifqi

Extraction de motifs graduels par corrélations d'ordres induits

Extracting gradual itemsets based on rank correlations

A. Laurent[1] M.-J. Lesot[2] M. Rifqi[2]

[1] LIRMM-Univ. Montpellier 2 - CNRS UMR 566 - 161 rue Ada, 34095 Montpellier
[2] LIP6 - UPMC - CNRS UMR 7606 - 4 place Jussieu, 75005 Paris

Résumé :

Les tendances graduelles de la forme *plus X est A, plus Y est B* expriment linguistiquement des informations sur les corrélations et co-variations des attributs. Dans cet article, nous présentons une étude comparative des formalisations qui ont été proposées, examinant leurs sémantiques et propriétés respectives. Nous proposons ensuite un algorithme qui combine les principes de plusieurs approches existantes pour extraire efficacement les motifs graduels fréquents et nous illustrons son utilisation sur une base de données réelle.

Mots-clés :

fouille de données, motifs graduels, comparaison d'ordres

Abstract:

Gradual tendencies of the form *the more X is A, the more Y is B* linguistically express information about correlation between attributes and their covariations. In this paper, we present a comparative study of the various formalisations that have been proposed, studying their respective semantics and properties. We then propose an algorithm that combines the principles of existing approaches to efficiently extract frequent gradual itemsets, illustrating its use on a real data set.

Keywords:

data mining, gradual itemsets, ranking comparisons

1 Introduction

L'extraction d'informations permettant de décrire ou de caractériser des données, leurs tendances ou leurs comportements exceptionnels, peut prendre de nombreuses formes, qui fournissent différents types de connaissances. Nous considérons ici les règles graduelles, c'est-à-dire des règles exprimées linguistiquement sous la forme "plus X est A, plus Y est B", comme par exemple "plus la vitesse est élevée, plus le danger est grand", qui établissent des liens entre les variations des attributs.

De telles règles ont d'abord été introduites comme un cas particulier de règles d'inférence dans un cadre de logique floue [2, 6, 8, 7] : elles sont interprétées comme une généralisation des règles d'association à des données floues, énonçant que la présence floue d'une modalité A implique, au sens logique, la présence floue d'une modalité B. Elles expriment donc des contraintes sur chaque donnée individuellement, imposant que les degrés d'appartenance aux modalités présentes dans la règle vérifient une implication floue, modélisée par une r-implication. Des opérateurs de s-implication permettent de modéliser des règles de certitude, du type "plus on se lève tard, plus le retard est certain" [6, 8].

Une autre approche des règles graduelles, que nous considérons ici, les interprète comme des tendances globales qui s'appliquent aux données dans leur ensemble : elles sont considérées comme imposant des corrélations aux variations des valeurs des attributs ou des degrés d'appartenance aux modalités [9, 1, 4, 5] : un accroissement de A doit aller de pair avec un accroissement de B. Différentes mesures de qualité pour implémenter ce principe général ont été proposées, associées à des sémantiques qui conduisent à l'identification de différents types de motifs graduels.

Dans cet article nous proposons une étude comparative de ces formalisations, discutant leurs sémantiques et leurs propriétés. Nous proposons ensuite de combiner plusieurs principes en une nouvelle méthode : nous replaçons la définition de motif graduel initialement proposée par [1] dans le contexte des mesures de com-

paraison d'ordres, et proposons une méthode d'extraction efficace inspirée de l'algorithme introduit par [5].

L'article est organisé de la façon suivante : la section 2 présente l'étude comparative des formalisations existantes des règles graduelles comme contraintes de covariations. La section 3 décrit l'interprétation proposée en termes de corrélation d'ordres induits basée sur le τ de Kendall et la section 4 présente l'algorithme d'extraction associé. Enfin, la section 5 décrit les résultats expérimentaux obtenus sur une base de données réelle.

2 Etude comparative des formalisations des règles graduelles

2.1 Notations et définitions

Dans cette section, nous rappelons brièvement les définitions classiques des items, motifs et règles graduels, en considérant successivement les cas des données numériques et des données floues, et en soulignant leurs différences.

Données numériques. On note $\mathcal{D}$ un ensemble de données numériques contenant n objets, décrits par p attributs. Un *item graduel* est défini comme un couple constitué d'un attribut A et d'un sens de variation noté $<$ ou $>$: l'item graduel $A^{>}$, resp. $A^{<}$, représente alors l'expression linguistique "plus A est élevé", resp. "moins A est élevé", et exprime une contrainte d'ordre croissant, resp. décroissant. On peut également considérer les sens de variation $\leq$ et $\geq$, qui signifient que les contraintes autorisent les cas d'égalité, c'est-à-dire des ex aequo, et non seulement des ordres stricts.

Un *motif graduel* $M = \{A_j^{*_j}, j = 1..k\}$, avec $\forall j = 1..k, *_j \in \{>, <\}$, est défini comme un ensemble d'items graduels, interprété comme leur conjonction : il impose une contrainte sur plusieurs attributs simultanément. Ainsi, le motif $A^{>}B^{<}$ est interprété comme "plus A est élevé et moins B est élevé".

Une *règle graduelle*, notée $M_1 \rightarrow M_2$, est alors définie comme un couple de motifs graduels qui satisfait une relation de causalité. Elle peut par exemple prendre la forme "plus la vitesse est élevée, plus le danger est grand", signifiant qu'une augmentation de la vitesse implique une augmentation du danger : elle rompt la symétrie du motif graduel dans lequel tous les items jouent un rôle identique.

Données floues. De nombreux travaux sur les règles graduelles considèrent des données dont les attributs (p. ex. vitesse) sont des variables linguistiques associées à des modalités floues (p. ex. *normale* et *élevée*). Les données sont alors décrites par des degrés d'appartenance indiquant à quel point leurs caractéristiques appartiennent à chaque modalité. Les valeurs d'attributs ne sont pas des ensembles flous, mais des degrés d'appartenance ; par abus de langage, nous désignons ces données comme floues.

Un item graduel flou est un triplet $(X, A, *)$ constitué d'un attribut X, une modalité A et un sens de variation $* \in \{>, <\}$. On l'interprète comme "plus (resp. moins) X est A", ou plus précisément "plus (resp. moins) le degré d'appartenance de X à A est élevé".

Cette seconde formulation permet de faire une analogie avec le cas non flou, en considérant que la règle ne s'applique pas dans l'univers de définition des attributs, mais dans l'univers des degrés d'appartenance, $[0, 1]$. Si on considère qu'on introduit un attribut par modalité, créant par exemple les attributs vitesseNormale et vitesseElevée, qui prend pour valeurs les degrés d'appartenance aux modalités correspondantes, on a équivalence formelle entre les motifs graduels pour données numériques et pour données floues, qui permet d'appliquer les mêmes méthodes d'extraction dans les deux cas. Dans la suite, nous utilisons la notation $A^{>}$ et $A^{<}$ pour les cas flous et non flous, sans les distinguer ; de plus, pour tout $x \in \mathcal{D}$, $A(x)$ désigne la valeur prise par l'attribut A pour l'objet x, ou le degré d'appartenance à la modalité correspondante de l'attribut correspondant dans le cas flou.

Il faut cependant souligner une différence sémantique entre les deux types de motifs, que nous illustrons en considérant l'exemple de l'item graduel "plus la vitesse est élevée". Dans un cas non flou, il exprime une contrainte d'ordre sur tout l'univers des vitesses, défini par exemple comme $[0, 200]$. Dans un cas flou, "élevé" est associé à une modalité floue définie pour la variable linguistique "vitesse", par exemple de noyau $[130, 200]$ et de support $[100, 200]$. La contrainte d'ordre s'applique alors aux degrés d'appartenance, dans l'univers $[0, 1]$. Elle ne fait donc intervenir que les vitesses supérieures à 100, restreignant les données qui supportent le motif et lui donne une interprétation plus "locale".

Une différence théorique entre données numériques et floues est également à noter : pour un motif graduel flou $M = \{A_j^{*j}, j = 1..k\}$, si tous les sens de variation sont identiques, on peut définir un degré d'appartenance au motif comme $M(x) = \top_{j=1..k}(A_j(x))$ où $\top$ désigne une t-norme, puisque le motif est interprété comme une conjonction des items qu'il contient. Dans le cas de données numériques, l'agrégation des valeurs de plusieurs attributs impliqués dans un motif graduel est potentiellement plus problématique, et sa sémantique doit être examinée en fonction des données considérées.

Les règles graduelles ont d'abord été introduites pour les données floues, dans un contexte de logique floue, basées sur des r-implications [2, 6, 8, 7]. Dans cet article, nous considérons des interprétations en termes de contraintes de covariations des attributs. Différentes approches ont été proposées, avec différentes sémantiques et mesures de support pour les motifs graduels, nous les examinons ci-dessous.

2.2 Interprétation comme covariation des valeurs d'attributs

Une première approche des règles graduelles comme contrainte de covariation de valeurs interprète une règle $A^> \rightarrow B^>$ comme une contrainte signifiant qu'une augmentation des valeurs de A est accompagnée d'une augmentation des valeurs de B [9].

Pour identifier de telles règles, il est proposé [9] d'appliquer une analyse de régression linéaire aux couples $(A(x), B(x))$. La validité de la règle est ensuite mesurée par la qualité de la régression, c'est-à-dire par le coefficient de corrélation linéaire ainsi que par la pente de la droite. Ainsi les attributs qui ne sont pas suffisamment corrélés sont rejetés, de même que ceux pour lesquels le degré d'appartenance à A reste constant alors que celui de B varie, ou inversement.

2.3 Interprétation comme règles d'association

Alors que la méthode précédente repose sur les valeurs numériques des attributs, d'autres approches considèrent la contrainte de covariation en termes de corrélation d'ordres : elles imposent que les classements des données induits par les attributs intervenant dans le motif soient identiques. Ainsi, dans le cas de motifs de taille 1, la règle graduelle "plus X est A, plus Y est B" est considérée comme valide si $\forall x, x' \in \mathcal{D}$, $A(x) < A(x')$ implique $B(x) < B(x')$; dans le cas de motifs tels que "plus X est A, moins Y est B", la contrainte impose que les ordres induits soient inversés.

Dans [1], il est proposé de formuler l'extraction de telles tendances par la découverte de règles d'association dans un ensemble approprié de transactions $\mathcal{D}'$ déduit de la base de données initiale $\mathcal{D}$: chaque paire d'objets de $\mathcal{D}$ est associée à une transaction t, qui possède alors un item A^* si la paire (x, x') associée vérifie la contrainte imposée par A^*, c'est-à-dire si $A(x) * A(x')$.

Une règle graduelle dans $\mathcal{D}$ est alors équivalente à une règle d'association classique extraite de $\mathcal{D}'$. Le support d'un motif graduel est donc défini comme le support de la règle d'association correspondante. Avec la définition précédente de la base de transactions, en notant n le nombre d'objets de $\mathcal{D}$, et pour un motif graduel

$M = \{A_j^{*j}, j = 1..k\}$ il s'écrit

$$S = \frac{|\{(x, x')/\forall j \ A_j(x) *_j A_j(x')\}|}{n(n-1)/2} \quad (1)$$

De même, la confiance des règles d'association conduit à définir la confiance pour les règles graduelles.

Dans [1], il est proposé un algorithme basé sur une discrétisation des valeurs observées, permettant de réduire le nombre d'individus à un nombre fixé de bins. Sa complexité reste cependant élevée et limite son application à des motifs de faible longueur.

2.4 Sous-ensembles compatibles

Dans [4, 5], une autre interprétation de la sémantique de la contrainte de covariation est considérée : elle propose d'identifier les sous-ensembles $\mathcal{D}^*$ de $\mathcal{D}$ qui peuvent être ordonnés de façon à ce que tous les couples de données de $\mathcal{D}^*$ vérifient les contraintes d'ordre. Ainsi, pour un motif $M = \{A_j^{*j}, j = 1..k\}$, $\mathcal{D}^* = \{x_1, ... x_m\} \subseteq \mathcal{D}$ doit être tel qu'il existe une permutation π telle que $\forall j \in [1, k], \forall l \in [1, m-1]$, $A_j(x_{\pi_l}) *_j A_j(x_{\pi_{l+1}})$. Le support du motif est alors défini comme $S = \max_{\mathcal{D}^* \in \mathcal{L}} |\mathcal{D}^*|/|\mathcal{D}|$, où $\mathcal{L}$ désigne l'ensemble de tous les sous-ensembles maximaux $\mathcal{D}^*$ compatibles avec le motif. La contrainte de maximalité impose qu'on ne peut ajouter d'objet à $\mathcal{D}^*$ sans perdre la propriété de compatibilité avec le motif graduel donné.

Une méthode par niveau basée sur des ensembles de conflit a été proposée pour identifier de tels motifs graduels [4] : elle consiste à éliminer à chaque niveau les données qui empêchent le plus grand nombre de données de satisfaire les contraintes considérées. Il s'agit d'une heuristique, car le choix d'une autre donnée à un niveau peut conduire à de meilleurs résultats au niveau suivant.

Dans [5] est proposée une méthode exacte basée sur les graphes de précédence, c'est-à-dire des graphes dont les nœuds sont les données et les arêtes représentent les relations de précédence. Le graphe est représenté par sa matrice d'adjacence, dans un format binaire : pour un motif $M = \{A_j^{*j}, j = 1..k\}$, le coefficient correspondant au couple de données (x, x') vaut 1 si $\forall j \in [1, k] \ A_j(x) *_j A_j(x')$, 0 sinon. Le support précédent peut alors être obtenu comme la longueur du chemin maximal dans le graphe.

La pertinence de cette approche vient de sa grande efficacité à générer les motifs graduels de taille $k+1$ à partir des motifs de taille k. En effet, si M est un motif généré à partir de deux motifs M' et M", sa matrice d'adjacence est obtenue par ET logique, élément par élément, des matrices associées à M' et M". De plus, des procédures d'élimination des lignes et des colonnes nulles de la matrice sont exploitées pour réduire la complexité spatiale de la méthode.

2.5 Rôle de l'amplitude des déviations

Au-delà des différences d'interprétation déjà soulignées, les méthodes précédentes diffèrent par leur traitement des amplitudes de déviation par rapport aux règles considérées, comme illustré sur la figure 1. Celle-ci représente deux ensembles de données décrits par deux attributs, pour lesquels l'objet o_2 est en contradiction avec le motif graduel $A^{>}B^{>}$, c'est-à-dire "plus A est élevé et plus B est élevé". Toutefois, pour l'ensemble de gauche, noté $\mathcal{D}_1$, la déviation à laquelle mène o_2 est moins importante que pour l'ensemble de droite noté $\mathcal{D}_2$.

Pour la définition par régression, cette différence est reflétée par une corrélation linéaire plus forte pour $\mathcal{D}_1$ que pour $\mathcal{D}_2$. Dans l'approche basée sur les règles d'association, le support est également plus petit pour $\mathcal{D}_2$ que pour $\mathcal{D}_1$: o_2 conduit à un nombre plus important de couples de données qui ne vérifient pas le motif graduel, à savoir (o_2, o_3), (o_2, o_4), (o_2, o_5) et (o_2, o_6), alors que pour $\mathcal{D}_1$, seul le couple (o_2, o_3) contredit le motif.

Au contraire, dans l'approche par sous-ensembles compatibles, $\mathcal{D}_1$ et $\mathcal{D}_2$ conduisent au

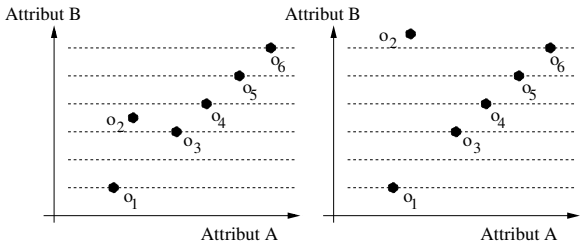

Figure 1 – Rôle de l'amplitude de déviation.

même support : dans les deux cas, il suffit de supprimer l'objet o_2 pour que les données restantes soient compatibles avec le motif.

3 Approche proposée : comparaison d'ordres

L'approche proposée vise à pouvoir être appliquée aux données numériques comme floues, prendre en compte les amplitudes de déviation et être associée à une extraction efficace.

3.1 Définition du support

La définition de support de l'équation (1) est présentée dans [1] dans un cadre de règle d'association, comme une généralisation du support classique à une base de transactions dérivée des données initiales. Toutefois, elle peut également être interprétée dans le cadre de la comparaison d'ordres. En considérant en effet que chaque attribut induit un ordre sur les données, ce support peut être vu comme quantifiant la ressemblance, ou le degré d'accord, entre ces ordres.

Comparaisons d'ordres. La notion de corrélation d'ordre a été largement étudiée par les statisticiens et plusieurs mesures ont été proposées : pour la comparaison de deux ordres, les mesures les plus utilisées sont la corrélation de Spearman et le τ de Kendall. Ce dernier est particulièrement intéressant dans le cadre des motifs graduels, car sa définition est directement liée à la définition de support de l'équation (1) : étant donné n objets à classer et σ_k, $k = 1, 2$ deux classements tels que $\sigma_k(x)$ donne le rang de l'objet x d'après σ_k, le τ de Kendall repose sur la définition de paires concordantes c'est-à-dire les paires d'objets (x, x') pour lesquelles les deux classements sont en accord, c'est-à-dire soit $\sigma_1(x) < \sigma_1(x')$ et $\sigma_2(x) < \sigma_2(x')$, soit $\sigma_1(x) > \sigma_1(x')$ et $\sigma_2(x) > \sigma_2(x')$. Le τ de Kendall est alors défini comme la proportion de paires non concordantes. Or la définition de support de l'équation (1) est égale à la proportion de paires concordantes.

Pour la comparaison de plus de deux classements, nécessaire pour des motifs de taille supérieure à 2, de nombreux critères ont également été proposés, avec l'objectif de répondre à des tests de significativité, pour déterminer si les différences entre les classements sont significatives [10]. On peut citer par exemple le coefficient W, le critère de Spearman moyen ou le test de Friedman, entre lesquels des relations statistiques ont également été établies [10].

Application à la définition de support. Les critères précédents, en dépit de leurs propriétés théoriques, ne peuvent être appliqués à l'évaluation des motifs graduels fréquents : ils ne possèdent pas de propriété de monotonie qui permette d'élaguer efficacement l'ensemble des motifs candidats quand on passe d'un niveau au suivant. Plus précisément, étant donné un ensemble de classements, l'ajout d'un nouveau classement peut conduire à une augmentation ou une diminution du coefficient W par exemple. Ceci signifie que même si un motif graduel est rejeté parce qu'ayant un coefficient W trop faible, il peut être nécessaire de considérer des motifs plus longs qui le contiennent, ce qui conduirait à des coûts de calcul trop élevés.

Au contraire, la comparaison d'ordres par le critère défini dans [1], même si elle ne permet pas de répondre à des tests de significativité, est anti-monotone : l'addition d'un nouvel item ne peut que diminuer la valeur du support.

La question est alors de calculer efficacement cette quantité. Nous proposons une approche classique par niveau, qui identifie les motifs de taille k pertinents à partir des motifs pertinents de taille $k-1$, la pertinence étant défi-

nie comme une valeur de support supérieure à un seuil donné par l'utilisateur. L'approche proposée est inspirée de la méthode basée sur le graphe de précédence, et en particulier la représentation par matrices binaires [5].

3.2 Liste de couples concordants

La valeur de support contient une information agrégée, qui résume la liste des paires concordantes à sa cardinalité. Elle ne permet pas de passer de deux motifs à un motif obtenu par fusion : en effet, on ne peut déterminer si une paire d'objets fixée est concordante pour les deux motifs, ou seulement l'un d'entre eux, et donc si elle reste concordante pour le motif joint. Aussi, il est nécessaire de conserver la liste des paires concordantes pour chaque motif.

De plus, pour prendre en compte le sens de variation associé aux motifs, il est nécessaire de considérer les couples d'objets et non les paires, en distinguant si le couple d'objets (x, x') ou le couple (x', x) est concordant : quand on considère des motifs réduits à un seul item, $A^{>}$ ou $A^{<}$, les paires suffisent, car les deux couples ont des statuts complémentaires. En effet, si l'un est concordant, l'autre est discordant. Dans le cas des motifs plus longs, par exemple $A^{>}B^{>}$, toutefois, il se peut que tous les deux soient discordants en raison d'incompatibilités avec des items différents au sein du motif considéré.

Il faut cependant noter que la longueur de cette liste de couples concordants doit être normalisée par le nombre total de *paires* possibles, $n(n-1)/2$, qui correspond à la valeur obtenue dans le cas de classements identiques.

3.3 Agrégation de listes de concordance

L'utilisation de telles listes de concordance est similaire à l'approche heuristique par ensemble de conflits [4], dans laquelle des listes de paires discordantes sont utilisées. La différence vient de ce que ces listes sont associées à chaque donnée dans [4] alors que nous proposons de les associer aux motifs. L'intérêt de ce changement de niveau est qu'il conduit à une méthode exacte pour généraliser les motifs de longueur $k-1$ aux motifs de longueur k.

En effet, la liste des couples concordants d'un motif M généré à partir des motifs M' et M'' est égale à l'intersection de leurs listes. Formellement, si M est généré à partir de M' et M'', il ne diffère de M' que par un item présent dans M'' (et réciproquement) : sans perte de généralité, si on note $M' = A_1^{*_1} \dots A_{k-1}^{*_{k-1}} B^{*_B}$ et $M'' = A_1^{*_1} \dots A_{k-1}^{*_{k-1}} C^{*_C}$, alors $M = A_1^{*_1} \dots A_{k-1}^{*_{k-1}} B^{*_B} C^{*_C}$. Aussi, un couple d'objets qui satisfait toutes les contraintes contenues dans M' et M'', c'est-à-dire exprimées par les items $A_j^{*_j}$, B et C, vérifie aussi tous les items présents dans M. Réciproquement, s'il vérifie tous les items de M, il satisfait aussi toutes les contraintes contenues dans M' et M''.

3.4 Représentation binaire

Le problème est alors de définir une méthode efficace pour stocker et traiter les listes de couples concordants. Pour cela, nous proposons d'utiliser une représentation binaire, comme proposée par [5] : une matrice de concordance binaire est définie pour chaque motif $\{A_j^{*_j}, j = 1..k\}$, telle que la valeur associée au couple (x, x') est 1 si $\forall j \in [1, k]\ A_j(x) *_j A_j(x')$, et 0 sinon.

D'une part cette représentation fournit une méthode efficace pour passer de motifs de longueur $k-1$ à des motifs de longueur k, puisque l'intersection des listes de concordance est équivalente à un ET logique appliqué aux matrices correspondantes. D'autre part, le support d'un motif est obtenu très simplement, comme la somme des éléments de la matrice, divisée par le nombre total de paires d'objets. Enfin, ces matrices deviennent rapidement creuses et leur taille peut être optimisée, en supprimant les lignes et colonnes qui ne contiennent que des 0.

4 Algorithme proposé

L'algorithme proposé suit donc le principe de l'algorithme Apriori, en modifiant l'étape

d'évaluation des motifs candidats qui utilise la méthode présentée dans la section précédente. Son déroulement est précisé dans le tableau 1.

L'efficacité de cet algorithme provient de plusieurs composantes : le calcul du support ne nécessite pas d'interroger la base de données, il est effectué à partir des informations obtenues au niveau précédent et cette information est manipulée efficacement, grâce à la représentation binaire des matrices de concordance.

La méthode proposée offre donc une implémentation efficace de la définition du support de motifs graduels introduite par [1], en l'interprétant dans le cadre de la comparaison d'ordres. D'un point de vue computationnel, l'algorithme bénéficie de l'efficacité de la représentation binaire [5], qu'il diminue puisque le coût du calcul de la somme de la matrice est inférieur à celui de la recherche du chemin le plus long. D'un point de vue sémantique, la méthode proposée permet de prendre en compte l'amplitude de la déviation des données qui ne vérifient pas les motifs graduels, en leur appliquant une pénalité variable.

5 Résultats expérimentaux

A titre d'illustration, nous présentons les résultats obtenus sur la base de données winequality-red [3] de UCI, qui décrit 1599 vins rouges selon 11 propriétés chimiques et leur qualité.

Le tableau 2 contient les motifs graduels dont le support est supérieur à 0.6, c'est-à-dire tels que 60% des paires de données vérifient les contraintes du motif. Les motifs obtenus reflètent des connaissances chimiques de base, qui doivent en effet être vérifiées dans les données. Ainsi, l'augmentation de l'acidité, fixe ou volatile, est associée à la diminution du pH. On peut supposer que le fait que leurs supports ne soient pas plus élevés est dû à la présence de valeurs égales dans les données, et donc de couples de données ex aequo. Ces derniers ne sont en effet pas comptés comme des paires concordantes qui servent de support au motif.

1. Initialisation ($k = 1$) : pour chaque attribut A, construire les matrices de concordance pour $A^>$ et $A^<$.
2. Génération des motifs de longueur $k + 1$: appliquer la procédure de génération d'Apriori, en calculant les matrices de concordance des motifs générés comme le ET logique des matrices associées aux motifs joints.
3. Evaluation des motifs candidats :
   (a) pour chaque candidat, calculer le support, comme la somme de la matrice de concordance divisée par $n(n-1)/2$.
   (b) éliminer les candidats dont le support est inférieur au seuil donné par l'utilisateur.
4. Itération des étapes (2) et (3) jusqu'à ce que la génération ne produise plus de nouveaux candidats.

Tableau 1 – Algorithme d'extraction de motifs graduels par comparaison d'ordres induits

(free SO_2 > total SO_2 >)	0.77
(fixed acidity> pH<)	0.73
(fixed acidity> citric acid>)	0.71
(volatile acidity> citric acid<)	0.69
(citric acid> pH<)	0.66

Tableau 2 – Motifs graduels de support supérieur à 0.6

(total SO_2 > quality<)	0.38
(citric acid> quality>)	0.38
(fixed acidity> quality>)	0.35
(free SO_2 > quality<)	0.33
(sulphates> alcohol> quality>)	0.32
(volatile acidity> citric acid< quality<)	0.32
(volatile acidity> alcohol< quality<)	0.32
(pH> quality<)	0.32
(residual sugar> quality>)	0.31
(volatile acidity> sulphates< quality<)	0.31

Tableau 3 – Motifs graduels contenant l'attribut "qualité" de support supérieur à 0.3

Le tableau 3 contient les motifs graduels de support supérieur à 0.3 qui contiennent l'attribut de qualité. Comme précédemment, les faibles valeurs de support sont dues à l'exclusion des valeurs ex aequo dans le calcul : la qualité est mesurée par un entier compris entre 3 et 8, de nombreuses valeurs égales sont donc rencontrées.

Le relâchement des contraintes par l'autorisation des ex aequo n'est pas pertinent en général, car il conduit à une forte augmentation des motifs identifiés comme pertinents. En effet, dans un cas extrême, deux attributs constants A et B, prenant une seule valeur pour toutes les données, conduisent aux quatre motifs $A^{\leq}B^{\leq}$, $A^{\geq}B^{\geq}$, $A^{\geq}B^{\leq}$ et $A^{\leq}B^{\geq}$, avec un support égal à 1. Ils n'apportent cependant pas d'information. De plus, ils bruitent les motifs graduels pertinents, en les quadruplant par l'ajout de ces quatre motifs.

6 Conclusion

Après une étude comparative des formalisations des motifs graduels, nous avons présenté un algorithme pour leur identification qui intègre des paradigmes complémentaires : une définition du support en termes de corrélation d'ordre, l'efficacité d'une implémentation par niveau, l'exploitation d'une représentation binaire des informations. Il peut être appliqué pour des données numériques comme pour des données floues et prend en compte l'amplitude de déviation par rapport aux motifs candidats.

Une première perspective concerne une comparaison plus précise, au-delà du niveau sémantique considéré ici, des motifs graduels extraits selon les différentes méthodes sur des bases de données réelles. Une autre perspective vise à étudier la prise en compte contextuelle des ex aequo, par exemple par le biais de pondération. Celle-ci pose des problèmes de définition des cas d'application, comme discuté précédemment, mais aussi des difficultés calculatoires : l'utilisation de poids exclut la représentation binaire des matrices de concordance. Une autre perspective, pour le traitement de données floues, concerne l'extraction automatique des modalités floues pertinentes, permettant de rechercher des motifs graduels locaux ou contextuels.

Références

[1] 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.

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

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4) :547–553, 2009.

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

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

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

[7] S. Galichet, D. Dubois, and H. Prade. Imprecise specification of ill-known functions using gradual rules. *IJAR*, 35(3) :205–222, 2004.

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

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

[10] M. Kendall and B. Babington Smith. The problem of m rankings. *Annals of mathematical statistics*, 10(3) :275–287, 1939.