OSACA: Découverte d'attributs symboliques ordinaux

Christophe Marsala, Anne Laurent, Marie-Jeanne Lesot, Maria Rifqi, Arnaud Castelltort

OSACA : Découverte d'attributs symboliques ordinaux

OSACA: Discovering Ordinal Categorical Attributes

C. Marsala[1] A. Laurent[2] M.-J. Lesot[1] M. Rifqi[3] A. Castelltort[2]
[1] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6,
LIP6, F-75005 Paris, France, {prénom.nom}@lip6.fr
[2] LIRMM, Université de Montpellier, CNRS, Montpellier, France, {prénom.nom}@umontpellier.fr
[3] LEMMA, Université Panthéon-Assas, Paris, France, maria.rifqi@u-paris2.fr

Résumé :
Les bases de données dites hétérogènes contiennent des données décrites par des attributs à la fois symboliques et numériques. Cet article propose une méthode, appelée OSACA, pour identifier, parmi les attributs symboliques, les attributs ordinaux, en exploitant les informations fournies par les attributs numériques. Pour ce faire, OSACA procède en trois étapes : des motifs graduels sont d'abord extraits des attributs numériques. Des filtres morphologiques sont ensuite appliqués aux attributs symboliques pour déterminer des ordres sur les valeurs catégorielles à partir de l'ordre induit par les motifs graduels. Enfin, une mesure d'entropie d'ordre permet d'évaluer la pertinence des ordres candidats.

Mots-clés :
Attributs ordinaux, motifs graduels, mesure d'entropie d'ordre, morphologie mathématique.

Abstract:
This paper proposes to exploit heterogeneous data, i.e. data described by both numerical and categorical features, so as to discover whether, based on information provided by the numerical attributes, some categorical attributes actually are ordinal ones. The proposed 3-step methodology OSACA, first extracts gradual patterns from the numerical attributes ; it then applies mathematical morphology tools to induce an associated order on the categorical attributes. The third step evaluates the quality of the candidate rankings through measures derived from the rank entropy discrimination.

Keywords:
Ordinal Attributes, Gradual Patterns, Rank Discrimination Measure, Mathematical Morphology.

1 Introduction

Les données hétérogènes sont décrites à la fois par des attributs numériques et catégoriels et soulèvent des questions concernant leur exploitation simultanée, c'est-à-dire la combinaison de l'information qu'ils fournissent. Ainsi, le clustering de telles données peut être abordé en utilisant des approches relationnelles basées sur des mesures de distance appropriées [3, 14] ; la classification peut être réalisée par des arbres de décision qui traitent efficacement ces deux types d'attributs [19].

Cet article propose une autre approche qui consiste à exploiter les informations fournies par les attributs numériques pour mieux comprendre celles fournies par les attributs catégoriels. Plus précisément, le but est de déterminer si certains des attributs catégoriels peuvent être considérés comme ordinaux, et identifier, dans ce cas, l'ordre associé.

Le tableau 1 fournit un exemple illustratif : les données sont décrites par deux attributs numériques (X et Y) et un attribut catégoriel (couleur). Pour ces données, les valeurs des attributs numériques peuvent conduire à considérer l'attribut catégoriel comme ordinal, selon l'ordre partiel : bleu $\prec$ jaune.

Il faut noter qu'il existe de nombreuses mesures de distance et similarité pour les données catégorielles (voir, par exemple, [13]), toutefois, celles-ci sont le plus souvent fixées a priori, avant l'apprentissage. L'objectif dans cet article est, d'une part, de déterminer un ordre, ce qui est moins contraignant que de définir une mesure de comparaison, et, d'autre part, d'extraire automatiquement cet ordre par apprentissage.

Dans le but d'identifier des ordres, éventuellement partiels, sur les attributs catégoriels à partir des attributs numériques, l'article propose une méthode originale appelée

OSACA, pour *Order Seeking Algorithm for Categorical Attributes*, qui combine trois outils : les motifs graduels, la morphologie mathématique et les mesures d'entropie d'ordre. Les motifs graduels sont des motifs de la forme *plus/moins* a_1, ..., *plus/moins* a_k où les a_i sont des attributs numériques. Dans l'approche proposée, ils sont extraits pour établir des ordres sur les objets de la base de données. Ces ordres sont ensuite traités par des outils morphologiques, plus précisément par des filtres alternés, pour induire des ordres candidats sur les attributs catégoriels. Enfin, les candidats sont évalués par une mesure dérivée de la mesure d'entropie d'ordre.

L'article est organisé comme suit : la section 2 rappelle des définitions des trois types d'outils mentionnés ; la section 3 décrit l'approche OSACA et la section 4 présente des résultats sur des données artificielles illustrant la pertinence d'OSACA. La section 5 conclut l'article et présente quelques perspectives.

2 Préliminaires

Cette section rappelle les concepts-clés de chacun des trois outils utilisés dans la méthode OSACA : les motifs graduels, la mesure d'entropie d'ordre et la morphologie mathématique.

Dans cet article, $\Omega = \{o_1, \ldots, o_n\}$ désigne un ensemble de n objets, décrits par un ensemble de $m + p$ attributs, union de l'ensemble $\mathcal{N}$, contenant m attributs numériques, et de l'ensemble $\mathcal{C}$, contenant p attributs catégoriels. La valeur de l'attribut a pour l'objet o est notée $o[a]$.

2.1 Motifs graduels

Les motifs graduels extraient des connaissances linguistiques à partir de données décrites par des attributs numériques. Ils s'expriment sous la forme *plus/moins* a_1, ..., *plus/moins* a_k où $a_i \in \mathcal{N}$, par exemple *plus le budget est élevé, plus le nombre de victoires en ligue des champions est grand*. Initialement introduits dans le formalisme de l'implication floue [6, 7, 10], ils ont ensuite été interprétés comme exprimant des contraintes sur les covariations d'attributs. Plusieurs interprétations de ces contraintes ont été proposées : régression [11], corrélation de l'ordre induit [2, 12] ou existence de sous-ensembles d'objets compatibles [4, 5]. Chaque interprétation est associée à une définition de support pour quantifier la validité des motifs graduels et à une méthode pour l'identification des motifs fréquents selon la définition de support.

Tableau 1 – Exemple illustratif

Id	X	Y	coul.	Id	X	Y	coul.
o_1	0	1	bleu	o_5	2.5	9.8	rouge
o_2	1.2	1.5	bleu	o_6	3.0	2.1	bleu
o_3	1.8	1.6	rouge	o_7	4.8	3.2	jaune
o_4	2.3	9.3	jaune	o_8	5.0	8.5	jaune

Cette section détaille l'approche exploitée par OSACA, basée sur l'identification de sous-ensembles d'objets compatibles [4, 5],

Definition 2.1. Un *item graduel* est une paire $(a, *)$ où $a \in \mathcal{N}$ est un attribut numérique et $* \in \{\uparrow, \downarrow\}$ un sens de variation.

Definition 2.2. Un *motif graduel* P de taille k est un ensemble de k items graduels $\{(a_1, *_1), \ldots, (a_k, *_k)\}$, interprété comme leur conjonction.

Par exemple, $(budget, \uparrow)$ est l'item graduel *plus le budget est élevé* ; $\{(X, \uparrow), (Y, \uparrow)\}$ est le motif graduel *plus X est élevé et plus Y est élevé*.

La question principale est alors l'évaluation de la qualité d'un motif graduel candidat par rapport à l'ensemble de données considéré. Dans les méthodes classiques de découverte de motifs fréquents et de règles d'association, le support est défini comme le nombre d'objets contenant le motif [1]. Sa transposition au cas graduel [4, 5] propose de classer les objets par rapport au motif considéré, selon la définition

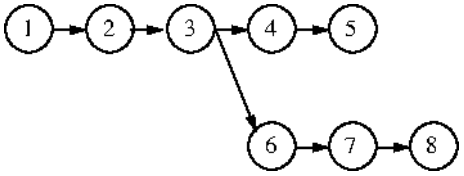

Figure 1 – Graphe de précédence pour le motif $\{(X,\uparrow),(Y,\uparrow)\}$ et les données issues du tableau 1. Les arcs de transitivité sont omis pour alléger le graphe.

d'une relation de précédence induite par un motif, qui définit un ordre partiel sur les objets.

Definition 2.3. Étant donné un ensemble Ω et un motif graduel $P = \{(a_1, *_1), \ldots, (a_k, *_k)\}$, la *relation de précédence induite par* P, notée $\prec_P$, est satisfaite pour le couple d'objets o et $o' \in \Omega$, $o \prec_P o'$, si et seulement si $\forall j \in [1,k]$:

— si $*_j = \uparrow$, $o[a_j] < o'[a_j]$
— si $*_j = \downarrow$, $o[a_j] > o'[a_j]$

Par exemple, avec les données du tableau 1, en considérant le motif $P = \{(X,\uparrow),(Y,\uparrow)\}$, on obtient $o_3 \prec_P o_4$. En effet, on a $o_3[X] = 1.8 < 2.3 = o_4[X]$ et $o_3[Y] = 1.6 < 9.3 = o_4[Y]$.

La relation de précédence conduit à la définition d'un graphe de précédence, où un nœud représente un objet et un arc, l'existence d'une relation de précédence, comme illustré sur la figure 1. Le graphe peut aussi être représenté de manière équivalente par sa matrice d'adjacence.

La relation de précédence conduit alors à la définition de chemin associé à un motif.

Definition 2.4. Étant donné un ensemble Ω et un motif graduel $P = \{(a_1, *_1), \ldots, (a_k, *_k)\}$, un *chemin* p de taille l associé à P est un ensemble ordonné $\langle o_{\rho_1}, \ldots, o_{\rho_l}\rangle$ contenant l objets de Ω tels que $\forall j \in [1, l-1]$, $o_{\rho_j} \prec_P o_{\rho_{j+1}}$.

On note $\mathcal{P}$ l'ensemble de tous les chemins associés à P.

Definition 2.5. Étant donné un ensemble Ω et un motif graduel P, le *support* de P dans Ω est défini par la longueur du chemin associé le plus long, relativement au nombre total d'objets :

$$supp(P) = \frac{1}{|\Omega|} \max_{p \in \mathcal{P}} (length(p)).$$

Ainsi, pour le motif $P = \{(X,\uparrow),(Y,\uparrow)\}$, les chemins sont $\langle o_1, o_2, o_3, o_4, o_5\rangle$ et $\langle o_1, o_2, o_3, o_6, o_7, o_8\rangle$ (ainsi que toutes les sous-chaînes de ces chaînes). Le support de P est donc $supp(P) = \frac{6}{8} = 0.75$.

Avec ce critère de qualité, on peut définir des motifs graduels d'intérêt, qui sont à la fois fréquents et maximaux.

Definition 2.6. Étant donné un ensemble Ω et un seuil de support minimal $minsup$, un motif graduel P est dit *fréquent* si $supp(P) \geq minsup$.

Definition 2.7. Étant donné un ensemble Ω et un seuil de support minimal $minsup$, un motif graduel fréquent P est dit *maximal* s'il n'existe aucun motif fréquent P' tel que $P \subset P'$.

L'algorithme GARE [5] est une méthode efficace pour extraire les motifs graduels fréquents maximaux selon ces définitions. En particulier, il permet d'éviter le parcours exhaustif de tous les sous-ensembles d'attributs numériques.

2.2 Mesure d'entropie d'ordre

Il existe de nombreux critères pour comparer des ordres entre eux et mesurer leur compatibilité ou leur cohérence, on peut citer à titre d'exemple, les indices de Kendall ou de Spearman. Dans cet article, nous proposons d'utiliser les mesures d'entropie d'ordre [16, 17] qui ont été initialement proposées pour des tâches de classification lorsque la classe à prédire est ordinale, et non catégorielle comme c'est habituellement le cas. Le but est alors de conserver, dans le classifieur appris, l'ordre sur la classe et de mettre en évidence une relation graduelle entre les attributs numériques et celle-ci.

Les mesures d'entropie d'ordre reposent sur le concept de *dominance* [8] : étant donné une relation d'ordre total $\preceq$ sur un ensemble d'objets Ω, l'ensemble dominant d'un objet $o \in \Omega$ est défini comme $[o]^{\preceq} = \{o' \in \Omega \ / \ o \preceq o'\}$.

La mesure d'entropie d'ordre de Shannon [9], exploitée par OSACA, correspond à une version modifiée de l'entropie de Shannon. Elle est obtenue en remplaçant la probabilité conditionnelle par une mesure de dominance :

Definition 2.8. Étant donné deux ordres totaux $\preceq_1$ et $\preceq_2$ définis sur un ensemble Ω d'objets, la mesure d'entropie d'ordre de $\preceq_1$ par rapport à $\preceq_2$ est définie comme :

$$H^*_S(\preceq_1 \mid \preceq_2) = -\frac{1}{|\Omega|} \sum_{o \in \Omega} \log_2 \left(\frac{|[o]^{\preceq_1} \cap [o]^{\preceq_2}|}{|[o]^{\preceq_2}|} \right)$$

Il faut souligner que, dans cette équation, l'intersection est nécessairement non vide car elle contient au moins l'objet o lui-même.

2.3 Induction d'ordre par morphologie mathématique

Les outils de la morphologie mathématique [20] permettent d'identifier des structures spatiales, en particulier en traitement d'images et en analyse fonctionnelle. Leur variante unidimensionnelle [15] s'applique à des séquences de symboles, appelées *mots*. Ainsi, les *filtres morphologiques alternés* ont été introduits et exploités pour construire automatiquement des partitions floues d'attributs numériques [15] ou pour enrichir des motifs graduels par des attributs catégoriels [18].

Étant donné une séquence de symboles, l'application d'un tel filtre fait ressortir des zones homogènes, c'est-à-dire des séquences d'un même symbole. Elles peuvent, par exemple, servir ensuite à définir des noyaux d'ensembles flous en retenant les séquences maximales [15].

Formellement, un filtre alterné est défini comme la composition d'opérations d'*ouverture* et de *fermeture*, elles-mêmes définies comme des compositions d'*érosion* et *dilatation*. Dans le cas unidimensionnel, une dilatation permet de fusionner les séquences d'un même symbole c séparées par "peu" d'occurrences de symboles différents, une érosion permet de supprimer les "petites" séquences d'un même symbole c, séparées par des symboles différents. Le symbole c est appelé *élément structurant*.

3 Méthode proposée : OSACA

Cette section présente l'approche OSACA qui repose sur trois étapes, basées sur les trois approches rappelées dans la section précédente, et décrites tour à tour dans les sections suivantes : la première exploite les attributs numériques dont elle extrait des motifs graduels, leurs supports et les chemins associés, afin d'induire des ordres candidats sur les objets. La seconde considère ces chemins du point de vue des attributs catégoriels et traite les mots qu'ils induisent en utilisant un filtre alterné, pour proposer un ordre candidat sur les valeurs des attributs catégoriels selon leurs séquences homogènes. La troisième étape évalue ces candidats et identifie les ordres pertinents, d'après une mesure de qualité prenant en compte à la fois la compatibilité de l'ordre des valeurs catégorielles avec l'ordre des motifs graduels et le nombre d'objets concernés par l'ordre évalué.

3.1 Identification de sous-ensembles d'objets compatibles

La première étape exploite les attributs numériques pour extraire des connaissances riches sous la forme de motifs graduels basés sur l'interprétation de co-variation : ils permettent de combiner de façon maximale les attributs induisant un ordre sur les objets.

OSACA exploite l'algorithme GARE [5], dont les principes sont rappelés dans la section 2.1 qui offre l'avantage additionnel d'extraire les chemins associés.

Étant donné l'ensemble d'objets Ω, cette étape

renvoie un ensemble de motifs graduels utilisant les attributs numériques $\mathcal{N}$, chacun d'entre eux étant associé à un ensemble de chemins $\mathcal{P}$.

3.2 Construction des ordres candidats

Dans la seconde étape, pour chaque motif graduel P extrait par GARE, tout chemin associé p induit un ordre candidat sur chaque attribut catégoriel $c \in \mathcal{C}$.

Étant donné un chemin $p = \langle o_1, \ldots, o_{|p|} \rangle$ de $\mathcal{P}$, afin de trouver l'ordre candidat sur c, on construit le mot constitué de la séquence correspondante des valeurs catégorielles : $w_p = \langle p_1[c], \ldots, o_{|p|}[c]$. Dans le cas de l'exemple du tableau 1, pour le motif graduel $P = \{(X, \uparrow), (Y, \uparrow)\}$ et le chemin de support maximal $p = \langle o_1, o_2, o_3, o_6, o_7, o_8 \rangle$, le mot obtenu est $w_p = \langle$bleu, bleu, rouge, bleu, jaune, jaune$\rangle$.

La mise en œuvre d'un filtre morphologique alterné permet alors de mettre en évidence des séquences homogènes pour chaque valeur catégorielle. Toutefois, dans l'approche originale de [15], la dilatation d'une séquence de c ne peut pas se faire au détriment d'autres symboles que l'élément structurant c. Dans le cadre considéré ici, il peut être nécessaire de supprimer de tels symboles, ce qui requiert un opérateur de dilatation spécifique : nous proposons un opérateur de *dilatation forte* permettant de mettre en œuvre cette suppression et appliquons le filtre alterné F qu'il induit.

Dans le cas de l'exemple courant, le mot w_p devient, après filtrage, $w'_p = \langle$bleu, bleu, $\cdot$, $\cdot$, jaune, jaune$\rangle$. Le symbole $\cdot$ désigne des valeurs catégorielles supprimées lors du filtrage, ici rouge et bleu qui sont trop isolées.

Le mot filtré $w'_p = F(w_p)$ permet enfin de définir un ordre partiel sur les valeurs de l'attribut considéré c : pour chaque valeur c_i, la séquence de taille maximale associée à c_i dans w' est retenue (la première en cas d'égalité) et l'ordre de ces séquences donne l'ordre sur les valeurs qui leur sont associées. Dans notre exemple, le mot w'_p induit l'ordre partiel bleu $\prec_p$ jaune.

Il faut noter que, dans certains cas, comme dans notre exemple, une valeur c_i peut ne pas être préservée après filtrage (ici rouge et bleu) : elle n'est alors pas concernée par l'ordre et le résultat $\prec_p$ est un ordre partiel . Dans le cas extrême où aucune valeur catégorielle n'est préservée, aucun ordre ne peut être induit : l'attribut catégoriel correspondant n'est pas ordinal, même partiellement.

3.3 Évaluation d'un ordre candidat

L'étape suivante vise à évaluer les ordres candidats $\prec_p$ ainsi identifiés, en combinant plusieurs critères de qualité.

D'une part, une notion de support est prise en compte, afin de quantifier la proportion de données concernées par l'ordre candidat. Ce support est défini comme la somme des tailles des séquences sur lesquelles l'ordre est construit, rapportée à la longueur du chemin, soit $\frac{|w'_p|}{|w_p|}$. Pour l'exemple considéré, le support de l'ordre partiel candidat bleu $\prec_p$ jaune est $4/6 = 0.666$.

D'autre part, la compatibilité entre l'ordre candidat sur les valeurs catégorielles et l'ordre du motif est considérée : elle est évaluée par la mesure d'entropie d'ordre de Shannon appliquée au sous-ensemble des données de Ω qui sont présentes dans le chemin considéré p. Les deux ordres comparés sont, d'une part, l'ordre $\prec_P$ associé au motif P considéré et, d'autre part, l'ordre $\prec_p$ induit pour l'attribut catégoriel par le chemin p associé à P.

Il faut souligner qu'il est possible que l'ordre induit $\prec_p$ ne s'applique pas à tous les points du chemin p, puisqu'il peut n'être que partiel, comme c'est par exemple le cas pour l'exemple considéré.

Aussi, nous introduisons, comme second critère de qualité d'un ordre candidat, *l'entropie d'ordre partiel de Shannon*, définie comme une extension de mesure d'entropie d'ordre (rap-

pelée dans la définition 2.8) : les ordres n'étant plus totaux, il faut noter que, contrairement à la mesure d'entropie d'ordre de Shannon, l'intersection $[o]^{\preceq_1} \cap [o]^{\preceq_2}$ peut être vide, ce qui peut conduire à une valeur indéterminée dans la somme. Dans ce cas, nous proposons que le terme correspondant soit considéré comme nul.

Pour l'exemple considéré, la mesure d'entropie d'ordre partiel de Shannon vaut $-\frac{1}{6}(\log_2(\frac{5}{6}) + \log_2(\frac{5}{5}) + 0 + \log_2(\frac{3}{3}) + \log_2(\frac{2}{2}) + \log_2(\frac{1}{1})) = 0.0659$. Cette valeur faible d'entropie indique que l'ordre partiel extrait est de qualité, conformément aux résultats attendus pour cet exemple.

4 Expérimentations

Des expériences ont été menées pour illustrer l'approche OSACA sur un ensemble de 100 données générées aléatoirement. Chaque donnée est décrite par deux attributs numériques X et Y et est associée à un attribut catégoriel (classe) parmi l'ensemble non ordonné {bleu, rouge, jaune, vert}. Les valeurs de X sont générées uniformément sur l'intervalle $[-20, 20]$, les valeurs de Y sont générées selon la loi $aX + b + \mathcal{U}$ où a et b sont des coefficients déterminant une relation graduelle linéaire entre X et Y et $\mathcal{U}$ définit un bruit uniforme.

4.1 Relation graduelle entre X et Y avec et sans ordre total sur la classe

Dans la première expérience, représentée sur la figure 2, les valeurs de classe ne sont pas corrélées avec les attributs numériques : il est attendu qu'aucun ordre ne soit identifié.

Comme il y a une relation graduelle entre les deux attributs X et Y, l'extraction de motifs graduels fournit plusieurs chemins pour $P = \{(X, \uparrow), (Y, \uparrow)\}$: elle fournit 16 chemins maximaux, chacun contenant 34 objets. Les chemins sont plutôt longs : ils contiennent un tiers des objets de la base de données, ce qui met en évidence la relation graduelle.

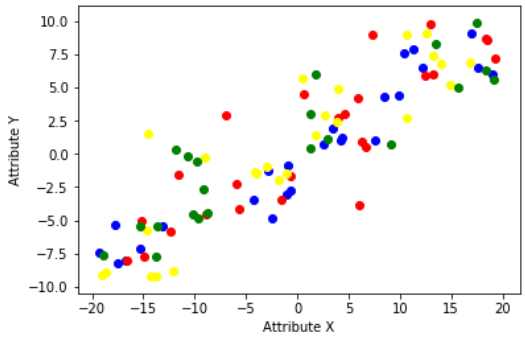

Figure 2 – Relation graduelle entre X et Y sans ordre sur l'atrtibut catégoriel.

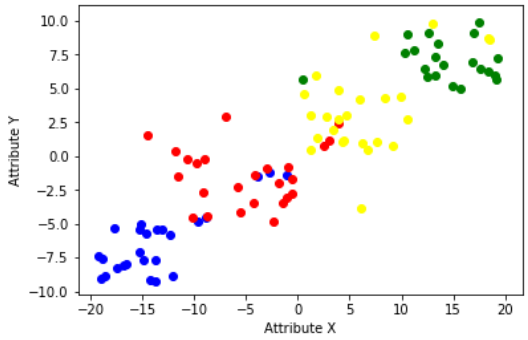

Figure 3 – Relation graduelle entre X et Y avec ordre total sur l'attribut catégoriel

Lors de l'application du filtre morphologique sur chacun de ces chemins, aucune séquence ne peut être trouvée : comme attendu, aucun ordre sur la classe ne peut être identifié.

Dans la deuxième expérience (figure 3) qui utilise les mêmes valeurs pour les attributs numériques X et Y, les quatre valeurs de classe sont totalement ordonnées : les plus petites valeurs des attributs numériques sont majoritairement associées à la classe bleu, les moyennes à rouge puis à jaune et les plus grandes à vert.

Le filtre morphologique appliqué à chacun des 16 chemins trouvés conduit à l'ordre attendu : bleu $\prec$ rouge $\prec$ jaune $\prec$ vert. Dans tous les cas, le support est 1 et l'entropie d'ordre de Shannon est très faible, comprise entre 0.02 et 0.03.

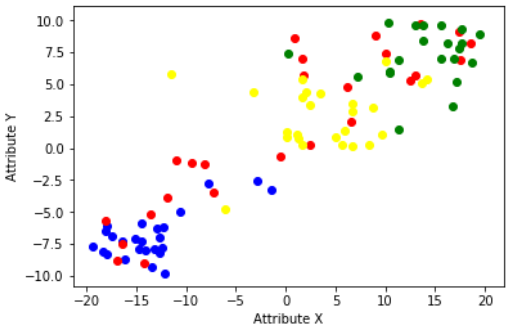

Figure 4 – Relation graduelle entre X et Y avec ordre partiel sur l'attribut catégoriel.

4.2 Relation graduelle entre X et Y avec ordre partiel sur la classe

Dans la troisième expérience (figure 4) qui utilise des valeurs d'attributs X et Y différentes, seulement trois des quatre classes (bleu, jaune et vert) sont générées avec un ordre sous-jacent, la classe rouge est uniformément répartie sur l'univers.

Comme dans l'expérience précédente, la relation graduelle entre les deux attributs numériques X et Y est mise en évidence par le fait que plusieurs chemins pour le motif graduel $\{(X, \uparrow), (Y, \uparrow)\}$ sont identifiés : avec ce jeu de données, 2160 chemins maximaux contenant chacun 24 objets sont trouvés. Ils contiennent le quart des objets de la base de données, ce qui met en évidence la relation graduelle existante entre X et Y.

Dans ce cas, l'ordre des valeurs de classe est identifié avec succès : lors de l'application du filtrage morphologique sur chacun de ces chemins, 1800 chemins parmi les 2160 mettent en évidence l'ordre bleu $\prec$ jaune $\prec$ vert, avec un support compris entre 0.75 et 0.875 et une entropie très petite comprise entre 0 et 0.008. Les autres chemins mettent en évidence l'ordre bleu $\prec$ vert avec un support compris entre 0.583 et 0.625. L'agrégation des différents résultats obtenus sur l'ensemble des chemins, ainsi que la question de leur stabilité ou de leur cohérence, reste une perspective à explorer.

Cette expérience illustre la capacité d'OSACA à identifier comme ordinal un attribut catégoriel même en présence d'un ordre partiel.

5 Conclusion et perspectives

Cet article propose l'approche OSACA qui permet de mettre en évidence parmi les attributs symboliques d'une base de données, ceux qui semblent être ordinaux. Quand un tel attribut est identifié, OSACA établit une hypothèse d'ordre dont la pertinence est testée grâce à une mesure d'entropie d'ordre originale. La méthode a été expérimentée sur des données synthétiques et se révèle prometteuse, ouvrant de nombreuses perspectives à ce premier travail.

Une des questions est de déterminer quels sont les motifs graduels sur lesquels la méthode doit s'appuyer : OSACA considère tous les motifs graduels dont le support est supérieur au seuil minimal défini en paramètre. Il pourrait être intéressant de ne considérer que les motifs maximaux au sens du nombre d'attributs ou du support : en raison de la propriété d'anti-monotonie, plus un motif est long, plus le support est faible. Une question liée concerne la stabilité et la cohérence des ordres ainsi construits à partir de plusieurs motifs, au delà du cas de plusieurs chemins pour un même motif évoqué précédemment, et leur évaluation respective par les mesures de qualité proposées.

Une autre perspective vise à étudier le choix des attributs symboliques pris en compte, pour en considérer un sous-ensemble voire un seul : OSACA considère tous les attributs symboliques, ce qui peut être coûteux et peu pertinent vis-à-vis des besoins de l'utilisateur.

Enfin, il serait intéressant d'enrichir l'évaluation de l'approche, en transposant les concepts de précision et rappel par exemple, afin de mesurer à quel point un attribut identifié par la méthode (avec son ordre) est pertinent (précision) ou au contraire à quel point tous les attributs symboliques ordinaux ont bien été identifiés (rappel).

Références

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD'93*, pages 207–216, 1993.

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

[3] Y.-M. Cheung and H. Jia. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, 46(8) :2228–2238, 2013.

[4] L. Di Jorio, A. Laurent, and M. Teisseire. Fast extraction of gradual association rules : A heuristic based method. In *5th Int. Conf. on Soft computing as transdisciplinary science and technology (CSTST'08)*, pages 205–210, 2008.

[5] L. Di Jorio, A. Laurent, and M. Teisseire. Mining frequent gradual itemsets from large databases. In *Int. Symp. on Intelligent Data Analysis (IDA'09)*, pages 297–308. Springer, 2009.

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

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

[8] S. Greco, B. Matarazzo, and R. Slowinski. Rough approximation by dominance relations. *Int. Journal of Intelligent Systems*, 17(2) :153–171, 2002.

[9] Q.-H. Hu, M.-Z. Guo, D.-R. Yu, and J.-F. Liu. Information entropy for ordinal classification. *Science China Inf. Sciences*, 53 :1188–1200, 2010.

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

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

[12] A. Laurent, M.-J. Lesot, and M. Rifqi. GRAANK : Exploiting rank correlations for extracting gradual dependencies. In *FQAS'09*, pages 382–393, 2009.

[13] M.-J. Lesot, M. Rifqi, and H. Benhadda. Similarity measures for binary and numerical data : a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(1) :63–84, 2008.

[14] C. Li and G. Biswas. Unsupervised learning with mixed numeric and nominal data. *IEEE TKDE*, 14(4) :673–690, 2002.

[15] C. Marsala and B. Bouchon-Meunier. Fuzzy partitioning using mathematical morphology in a learning scheme. In *Fuzz-IEEE'96*, pages 1512–1517, 1996.

[16] C. Marsala and D. Petturiti. Hierarchical model for rank discrimination measures. In *ECSQARU'13*, pages 412–423, 2013.

[17] C. Marsala and D. Petturiti. Rank discrimination measures for enforcing monotonicity in decision tree induction. *Information Sciences*, 291 :143–171, 2015.

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

[19] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1) :86–106, 1986.

[20] J. Serra. Introduction to mathematical morphology. *Comp. vision, graphics, and image process.*, 35(3) :283–305, 1986.