Quand les sous-groupes rencontrent les graduels : découverte de sous-groupes identifiant des corrélations exceptionnelles

Mohamed-Ali Hammal*, Céline Robardet*
Marc Plantevit**

*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621
**Université de Lyon, CNRS, Université Lyon 1, LIRIS UMR5205, F-69622, France
prénom.nom@liris.cnrs.fr

Résumé. La découverte de sous-groupes permet d'identifier des ensembles d'objets définis en intention qui sont intéressants vis-à-vis d'une mesure de qualité impliquant un ou plusieurs attributs cibles (par exemple motifs discriminants pour une variable de classe). Dans cet article nous proposons une approche pour un nombre quelconque (≥ 2) d'attributs cibles numériques. Pour cela, nous nous appuyons sur l'exploration conjointe de motifs graduels identifiant des corrélations de rang et de sous-groupes afin d'identifier des contextes pour lesquels les corrélations décrites par les motifs graduels sont exceptionnellement fortes par rapport au reste des données. Nous présentons un algorithme d'énumération s'appuyant sur des propriétés d'élagage avec des bornes supérieures. Une étude empirique sur plusieurs jeux de données démontre la pertinence et l'efficacité de notre méthode.

1 Introduction

Parmi les différentes techniques d'analyse exploratoire de données, la découverte de sous-groupes (Klösgen (1996)) vise à identifier des régions dans les données qui se détachent par rapport à une cible. Le principe est d'identifier des ensembles d'objets définis en intention qui sont fortement associés à certaines valeurs de la cible. Dans cet article nous proposons de généraliser cette approche au cas où l'on a plusieurs attributs cibles numériques. On cherche alors à la fois un sous-groupe d'objets défini par une conjonction de restrictions sur un ensemble d'attributs descriptifs et un sous-ensemble d'attributs cibles dont les valeurs sont fortement correlées sur cet ensemble. L'exploration conjointe de l'espace des descriptions et de l'espace des cibles permet de rechercher des corrélations pouvant être expliquées par d'autres variables descriptives de manière complètement non-supervisé.

Pour cela, nous introduisons le problème de **découverte de sous-groupes corrélés sur les rangs** basé sur l'exploration conjointe de motifs graduels identifiant des corrélations de rang et de sous-groupes afin d'identifier des contextes pour lesquels les corrélations sont exceptionnellement fortes par rapport au reste des données. Les motifs recherchés sont composés d'un ensemble D de conditions sur les attributs descriptifs, qu'ils soient numériques ou nominaux, et de C, un modèle de corrélation de rang sur des attributs numériques qui capture des corrélations de rang (positives ou négatives) basées sur une généralisation de τ de Kendall.

Nous présentons un algorithme d'énumération s'appuyant sur des propriétés d'élagage avec des bornes supérieures. Une étude empirique sur plusieurs jeux de données démontre la pertinence et l'efficacité de notre méthode.

2 Travaux connexes

Il existe plusieurs travaux visant à la découverte de motifs à forte co-variations entre des attributs numériques ou ordinaux. De telles approches sont connues sous plusieurs vocables : itemsets corrélés par les rangs (Calders et al. (2006)), dépendances graduelles (Hüllermeier (2002)), itemsets graduels (Do et al. (2010, 2015)) ou motifs de co-variation (Prado et al. (2013)). La plupart de ces approches considèrent la fouille de motifs pour lesquels la corrélation des attributs numériques est supérieure à un seuil défini par l'utilisateur. Par exemple, dans un article fondateur, Calders et al. (2006) proposent un processus d'extraction de motifs sous contrainte de correlation. Ils utilisent la mesure de corrélation τ de Kendall et proposent un processus d'élagage permettant la fouille de ce type de motifs. Cette approche a été généralisée dans Prado et al. (2013) pour découvrir des corrélations positives ou négatives entre un nombre quelconque (≥ 2) d'attributs. On peut noter que Calders et al. (2006) et Prado et al. (2013) introduisent des mesures supplémentaires qui prennent en compte la distribution de paires d'objets qui supportent les motifs : dans Calders et al. (2006), ces mesures caractérisent un seul attribut catégoriel avec des motifs corrélés sur les rangs, alors que Prado et al. (2013) introduit le concept de motifs émergents selon un seul attribut numérique ou un graphe. Ces deux approches ne considèrent néanmoins qu'une seule cible et ne visent pas à trouver des motifs exceptionnels en fonction de cibles multiples. Do et al. (Do et al. (2010, 2015)) utilisent une mesure de support basée sur la longueur du chemin le plus long entre les objets ordonnés par les attributs numériques. Cette mesure a plusieurs inconvénients, que ce soit au niveau calculatoire ou sémantique. Plus récemment, Downar et Duivesteijn (2017) ont proposé une approche pour trouver des sous-groupes dans le cas où deux cibles numériques interagissent de manière inhabituelle. L'interaction entre les deux cibles est modélisée par plusieurs mesures de corrélation (par exemple, le coefficient de corrélation de Pearson, la corrélation de rang τ de Kendall). La principale limitation de ce travail est que les deux attributs numériques cibles doivent être spécifiés apriori et l'approche ne fonctionne pas avec un ensemble arbitraire d'attributs numériques.

3 Sous-groupes corrélés sur les rangs

Un sous-groupe corrélé sur les rangs est un ensemble d'attributs fortement corrélés sur un ensemble d'objets défini en intention. Cet ensemble d'objets est identifié par une description, c'est-à-dire une conjonction de conditions sur des attributs descriptifs (par opposition aux attributs cibles). Plus formellement, un tel motif est composé de deux parties : un ensemble d'attributs corrélés positivement ou négativement, et une conjonction de restrictions sur certains attributs descriptifs. Les objets qui satisferont la description constituent le sous-groupe de données sur lequel les corrélations sont évaluées.

Avant d'introduire formellement le langage de motifs qui nous intéresse, établissons une notation. Dans ce qui suit, un ensemble de données est noté $\mathbb{D} = (\mathcal{O}, \mathcal{C}, \mathcal{R})$ où $\mathcal{O}$ est un

ensemble de n objets, $\mathcal{C}$ un ensemble d'attributs numériques qui associe à chaque objet une valeur réelle ($\forall c \in \mathcal{C},\ c\ :\ \mathcal{O} \rightarrow \mathbb{R}$). Des corrélations sont recherchées parmi les attributs de cet ensemble. $\mathcal{R}$ est un ensemble d'attributs qui peuvent être soit numériques soit catégoriels et dont la restriction de leurs domaines de valeurs identifie les sous-groupes.

3.1 Evaluer la corrélation d'un ensemble d'attributs

Les mesures de corrélation évaluent la force de l'association entre deux attributs ainsi que la direction de la relation. Trois types de corrélations sont utilisés en statistique : la corrélation de Pearson, le τ de Kendall et les corrélations de rang de Spearman. La corrélation de Pearson est la mesure la plus utilisée, mais elle nécessite des attributs continus, et pas seulement ordinaux. De plus, elle est basée sur des hypothèses fortes (les deux attributs doivent être distribués selon une loi normale, être corrélés avec relation linéaire et homoscédastique) qui ne sont généralement pas satisfaites dans la pratique. Des mesures de corrélation de rang (par exemple, la mesure de corrélation de rang τ de Kendall, ou de rang de Spearman) sont mieux adaptées car elles ne reposent pas sur les hypothèses mentionnées ci-dessus.

Contrairement au coefficient de Spearman, la mesure τ de Kendall est facile à interpréter et peut être facilement utilisée pour la fouille de motifs (Calders et al. (2006)).

Définition 1 (Motif corrélé sur le rang)**.** *Un motif corrélé sur les rangs C est un ensemble d'au moins deux attributs signés de $\mathcal{C}$ noté $C = \{(a, s) \mid a \in \mathcal{C}$ et $s \in \{-, +\}\}$ avec, par convention, le signe du premier attribut dans l'ordre canonique à $+$. Etant donné un ensemble d'objets $O \subseteq \mathcal{O}$, l'ensemble des paires concordantes d'objets O avec un motif C est défini par $\eta(C, O) = \{(o_i, o_j) \in O \times O \mid \nu_C(o_i, o_j)\}$ avec*

$$\nu_C(o_i, o_j) \equiv \bigwedge_{(a,s)\in C} (a(o_i) <_s a(o_j))$$

et $<_s$ est la relation binaire classique sur $\mathbb{R}$: $<$ quand $s = +$, et $>$ quand $s = -$. La mesure de corrélation τ de Kendall généralisée à un nombre quelconque d'attributs est alors :

$$\mathcal{T}(C, O) = \frac{|\eta(C, O)|}{N(O)} \text{ avec } N(O) = \left(\begin{array}{c} |O| \\ 2 \end{array} \right). \quad (1)$$

3.2 Corrélations de rang contextualisées

L'objectif du processus de fouille proposé est de trouver des sous-groupes d'objets O, définis en intention, pour lesquels la valeur de $\tau(C, O)$ sur les attributs C est plus forte que sur l'ensemble de tous les objets $\tau(C, \mathcal{O})$. Ces sous-groupes sont définis au moyen de conjonctions de restrictions sur les attributs de $\mathcal{R}$:

Définition 2 (Sous-groupe et support)**.** *Un sous-groupe d'objets est défini en intention par la description $D = \langle f_1, \ldots, f_{|\mathcal{R}|}\rangle$ avec chaque f_ℓ est une restriction sur le domaine de valeurs de l'attributs $d_\ell \in \mathcal{R}$. En fonction du type de d_ℓ, la restriction f_ℓ est définie par :*

- *$f_\ell = \{v\}$ avec $v \in$ **Dom**(d_ℓ), ou $f_\ell =$ **Dom**(d_ℓ), si d_ℓ est nominal*
- *$f_\ell = [v, w]$ avec $v, w \in$ **Dom**(d_ℓ) et $v < w$, si d_ℓ est numérique.*

L'ensemble des objets $\mathcal{O}$ qui vérifient D est appelé support de la description $D = \langle d_1, \ldots, d_{|\mathcal{R}|} \rangle$:

$$\sigma(D) = \{o_i \in \mathcal{O} \mid d_\ell(o_i) \in f_\ell,\ \forall \ell = 1 \ldots |\mathcal{R}|\}$$

Nous avons maintenant tous les ingrédients pour définir les sous-groupes corrélés.

Définition 3 (Sous-groupes corrélés)**.** *Un sous-groupe corrélé est une paire (C, D) avec C un motif de corrélation sur $\mathcal{C}$ et D une description sur $\mathcal{R}$ qui définit un sous-groupe d'objets en intention. La corrélation de C sur $\sigma(D)$ est mesurée par $\mathcal{T}(C, \sigma(D))$.*

Exemple 1 L'intérêt de ce modèle est illustré sur l'exemple du Tableau 1. Ces données décrivent un ensemble de baux commerciaux décrits par la date de début et de fin de bail, la localisation GPS du commerce, et son type. Les attributs de la cible décrivent l'environnement géographique du commerce sur la durée du bail, c'est-à-dire le nombre de commerces de chaque type se trouvant dans un rayon de 300m (pharmacie, boulangerie, boucherie). Un attribut supplémentaire indique la durée du bail du magasin (durée de vie). Sur cet exemple, le motif $C = Duree\_De\_Vie^+, Boulangerie^-, Boucherie^+$ est fortement corrélé sur le sous-groupe décrit par $D = x \in [1,3]$, $y \in [1,3]$, $cat \in \{Boulangerie\}$: on a $\tau(C, \sigma(D)) = 6/6$ alors que sur l'ensemble de tous les objets, $\tau(C, \mathcal{O}) = 6/21$. Ce motif indique que les magasins de boulangerie dans la région de coordonnées $[1,3] \times [1,3]$ durent d'autant plus longtemps qu'il y a peu de boulangeries mais beaucoup de boucheries dans leur voisinage.

Attributs descriptifs						Attributs cibles			
Id	DateDébut	DateFin	x	y	type	DuréeDeVie	pharmacie	boulangerie	boucherie
o1	1991	2000	1	3	boulangerie	10	5	7	1
o2	2000	2013	3	3	boulangerie	13	7	5	3
o3	1975	1992	2	1	boulangerie	18	3	2	7
o4	1986	2005	2	3	boulangerie	20	9	1	9
o5	1999	2008	2	3	Pharmacie	10	7	2	2
o6	1995	2014	5	3	boucherie	20	8	3	1
o7	1980	1999	4	4	boulangerie	20	6	3	1

TAB. 1 – *Exemple de données et de motif.*

Certains sous-groupes corrélés peuvent être considérés comme équivalents car ils partagent le même support. Ces motifs appartenant à une même classe d'équivalence peuvent être retirés par un opérateur de fermeture.

Définition 4 (Les opérateurs de fermeture)**.** *Suivant le formalisme de l'analyse formelle de concepts (Wille (1982)), on définit deux fonctions H and M qui permettent d'associer à un sous-groupe corrélé l'ensemble des paires d'objets qui le supportent et réciproquement :*

1. *$H(C, D) \equiv \eta(C, \sigma(D)) = \{(o_i, o_j) \in \sigma(D) \times \sigma(D) \mid \nu_C(o_i, o_j)\}$, comme défini ci-dessus (Définition 1).*
2. *Etant donné un ensemble de paires d'objets $X \subseteq \mathcal{O} \times \mathcal{O}$, $M(X)$ est le sous-groupe corrélé (C', D') défini par*
   - *$C' = \{(a, s) \in \mathcal{C} \times \{+, -\} \mid \forall (o_i, o_j) \in X,\ a(o_i) <_s a(o_j)\}$*
   - *$D' = \langle f'_1, \ldots, f'_{|\mathcal{R}|} \rangle$ avec*
     - *$f'_\ell = \begin{cases} v,\ Si\ \forall (o_j, o_k) \in X,\ d_\ell(o_j) = d_\ell(o_k) = v \\ \textbf{\textit{Dom}}(d_\ell)\ Sinon \end{cases}$ si d_ℓ est catégoriel,*

— $f'_\ell = [v, w]$ *avec* $\begin{array}{l} v = \min_{(o_i,o_j)\in X\cup(o_j,o_i)\in X} d_\ell(o_i) \\ w = \max_{(o_i,o_j)\in X\cup(o_j,o_i)\in X} d_\ell(o_j) \end{array}$ *si* d_ℓ *est numérique.*

Le couple $(H(C,D), M(X))$ *forme un concept formel.*

Les sous-groupes corrélés et fermés qui capturent le mieux les corrélations locales dans les données doivent avoir une valeur de corrélation élevée par rapport à ce qui est observé sur l'ensemble des données. Une mesure appropriée de ce phénomène est la précision relative pondérée (WRAcc) (Lavrač et al. (1999)). Cette mesure prend en compte l'accroissement de la précision par rapport à la corrélation par défaut, c'est-à-dire la corrélation sur l'ensemble de tous les objets.

Définition 5 (WRAcc)**.** *Le caractère exceptionnel d'un sous-groupe corrélé* (C, D) *est évalué en utilisant la mesure Wracc, définie comme suit :*

$$\textbf{\textit{WRAcc}}(C,D) \quad = \quad \frac{N(\sigma(D))}{N(\mathcal{O})}\left(\mathcal{T}(C,\sigma(D)) - \mathcal{T}(C,\mathcal{O})\right) \tag{2}$$

Notre tâche de fouille peut maintenant être entièrement exprimée comme le problème suivant :

Problème 1 (Fouille de sous-groupes corrélés et fermés)**.** *Sois* $\mathcal{S}$ *la collection de sous-groupes corrélés et fermés définie comme :*

$$\begin{array}{ll} \forall\,(C,D)\in\mathcal{S} & \\ & \textbf{\textit{Closure}}(C,D) = (C,D) \\ & \sigma(D) \geq \alpha \\ & \mathcal{T}(C,\sigma(D)) \geq \beta \\ & \textbf{\textit{WRAcc}}(C,D) \geq 0 \end{array}$$

Nous voulons également un ensemble concis de motifs inattendus qui maximisent la mesure WRAcc. Cependant, il est bien connu (Xin et al. (2006)) qu'en général ces motifs sont très redondants, certains étant une petite variation des autres. Nous proposons de limiter cette redondance à l'aide de l'appproche suivante :

Problème 2 (Fouille des top-k sous-groupes fermés, corrélés, exceptionnels et diversifiés)**.** *Sois* $\mathcal{K}$ *un sous-ensemble de* $\mathcal{S}$ *contenant les top-*k *sous-groupes fermés et corrélés par rapport à la mesure* ***WRAcc*** *et qui sont aussi diversifés. La diversité entre deux motifs est évaluée par la mesure de Jaccard, définie comme suit :*

$$\textbf{\textit{Jaccard}}((C,D),(C',D')) \quad = \quad \frac{|\eta(C,\sigma(D)) \bigcap \eta(C',\sigma(D'))|}{|\eta(C,\sigma(D)) \bigcup \eta(C',\sigma(D'))|} \tag{3}$$

Etant donné un seuil δ*, l'ensemble* $\mathcal{K}$ *des* k *sous-groupes les plus diversifiés, est défini par :*

1. $\forall (C,D),\ (C',D') \in \mathcal{K}^2$, ***Jaccard*** $((C,D),(C',D')) \leq \delta$
2. $\forall (C,D) \in \mathcal{K}$ *and* $\forall (C',D') \in \mathcal{S}$ *tel que* ***Jaccard*** $((C,D),(C',D')) > \delta$*, on a* ***WRAcc*** $(C,D) \geq$ ***WRAcc*** (C',D').
3. $|\mathcal{K}| = k$

Le point 2 est très difficile à garantir dans un processus incrémental. Cela est dû à la non-transitivitée de la mesure de similarité. En effet, si un sous-groupe corrélé (C, D) est exclu de $\mathcal{K}$ par un motif similaire (C', D') de meilleure qualité, il n'y a aucune garantie que d'autres sous-groupes corrélés exclus par similitude avec (C, D) soient également similaires à (C', D'). Nous relaxons donc cette condition de la manière suivante :

$$\forall (C, D) \in \mathcal{K},\ \exists (C', D') \in \mathcal{S} \text{ tel que } \textbf{\textit{Jaccard}}\left((C, D), (C', D')\right) > \delta$$
$$\text{et } \textbf{\textit{WRAcc}}(C, D) \geq \textbf{\textit{WRAcc}}(C', D')$$

4 Algorithme

Nous énumérons récursivement les sous-groupes corrélés par une recherche en profondeur DFS à l'aide de l'algorithme LoCoM (voir Algorithme 1). Étant donné le motif (C, D) actuellement exploré, l'algorithme retourne toutes ses spécialisations qui sont des sous-groupes corrélés exceptionnels. Pour le premier appel, le motif (C, D) est initialisé à $(\emptyset, M(\mathcal{O}))$. L'ordre de spécialisation considéré est $\preceq$, l'ordre partiel défini par : $(C, D) \preceq (C', D')$ si et seulement si $C \subseteq C'$ et, pour $D = \langle f_1, \ldots, f_{|\mathcal{R}|} \rangle$ et $D' = \langle f'_1, \ldots, f'_{|\mathcal{R}|} \rangle$, $f'_\ell \subseteq f_\ell$, $\forall \ell = 1 \ldots |\mathcal{R}|$. De plus, pour éviter de générer des motifs plusieurs fois, nous utilisons des ordres arbitraires, $\ll_{\mathcal{C}}$ sur $\mathcal{C}$ et $\ll_{\mathcal{R}}$ sur $\mathcal{R}$. L'ordre canonique entre les motifs est donc défini par $\ll$ avec :

$$\begin{aligned}(C, D) \ll (C', D') \quad \Leftrightarrow \quad & (C, D) \preceq (C', D') \\ & et\ \forall a \in C, \forall a' \in C' \setminus C, \quad a \ll_{\mathcal{C}} a' \\ & et : \text{argmax}_\ell f_\ell \neq \textbf{Dom}(d_\ell) \ll_{\mathcal{R}} \text{argmin}_\ell f'_\ell \neq \textbf{Dom}(d_\ell) \neq f_\ell\end{aligned}$$

Si X n'est pas vide (lignes 4 à 37), les sous-groupes corrélés actuels sont spécialisés soit en ajoutant un attribut signé en C, soit en réduisant la valeur du domaine d'un attribut dans $\mathcal{R}$. Si cet attribut est catégorique, son domaine est limité à une seule valeur (ligne 17). S'il est numérique, deux sous-intervalles peuvent être générés : un réduit d'une seule valeur sur la gauche (ligne 25) et l'autre à droite (ligne 32). Pour éviter de générer deux fois le même intervalle, la réduction sur la droite n'est autorisée que lorsqu'aucune réduction sur la gauche précédente ne s'est produite (Kaytoue et al. (2011)). La fonction de fermeture est utilisée pour faire *des sauts* dans le processus d'énumération, et obtenir directement le motif le plus spécifique couvrant les mêmes paires d'objets.

La fonction **Propager** $(X, (C_c, D_c))$ est utilisée pour éliminer rapidement les candidats peu prometteurs de X. Trois techniques d'élagage sont utilisées pour arrêter le processus d'énumération (ligne 6) : l'anti-monotonicité de la mesure support $\sigma(D)$, et deux bornes supérieures, une sur le τ de Kendall, et l'autre sur la mesure WRAcc.

Algorithme 1 : LoCoM((C, D), X, gauche)

Entrées : (C, D) le motif en cours de construction, X l'ensemble des couples attributs-valeurs de $\mathcal{C} \cup \mathcal{R}$, à énumérer. *left* : Un tabeau de $|\mathcal{R}_n|$ valeurs booléennes indiquant si les intervalles de l'attribut numérique correspondant ont été réduits sur le côté gauche.

Il y a aussi des variables globales :

— $\beta,\ \alpha,\ \delta$: les seuils utilisés pour les contraintes

— minWRAcc : la valeur WRAcc minimale des k motifs

Sorties : $\mathcal{K}$, Liste des top-k motifs diversifiés actuels.

si $X = \emptyset$ **alors**
  si $\mathcal{T}(C, \sigma(D)) \geq \beta$ *et* $\boldsymbol{WRAcc}(C, D) \geq minWRAcc$ **alors**
    $minWRAcc \leftarrow$ TopKDiv($\mathcal{K}$, (C, D))
sinon
  si $(UB_{\tau}(C, D) \geq \beta)$ *et* $(\sigma(S_D) \geq \alpha)$
  et $(UB_{\boldsymbol{WRAcc}}(C, D) \geq minWRAcc)$ **alors**
    Sois $(a, v) \in X$
    si $a \in \mathcal{C}$ **alors**
      $C' \leftarrow C \cup \{(a, v)\}$
      $(C_c, D_c) \leftarrow$ **Closure**(C', D)
      si $(C', D) \ll (C_c, D_c)$ **alors**
        LoCoM((C_c, D_c), **Propager**$(X, (C_c, D_c))$, $gauche$)
      LoCoM((C, D), $X \setminus \{(a, v)\}$, $gauche$)
    sinon
      //$a \in \mathcal{R}$
      si *(a est le $i$ème attribut symbolique de $\mathcal{R}$)* **alors**
        $D' = \langle f_1, \ldots, f_{i-1}, \{v\}, f_{i+1}, \ldots, f_{|\mathcal{R}|} \rangle$
        $(C_c, D_c) \leftarrow$ **Closure**(C, D')
        si $(C, D') \ll (C_c, D_c)$ **alors**
          LoCoM((C_c, D_c), **Propager**$(X, (C_c, D_c))$, $gauche$)
        LoCoM((C, D), $X \setminus \{(a, v)\}$, $gauche$)
      sinon
        // a est le $i$ème attribut numérique de $\mathcal{R}$
        $[x, y] \leftarrow v$
        $D' = \langle f_1, \ldots, f_{i-1}, [x + 1, y], f_{i+1}, \ldots, f_{|\mathcal{R}|} \rangle$
        $(C_c, D_c) \leftarrow$ **Closure**(C, D')
        si $(C, D') \ll (C_c, D_c)$ **alors**
          $gauche[i]' \leftarrow true$
          LoCoM((C_c, D_c), **Propager**$(X, (C_c, D_c))$, $gauche'$)
        LoCoM((C, D), $X \setminus \{(a, v)\}$, $gauche$)
        si *gauche[i]=faux* **alors**
          $D' = \langle f_1, \ldots, f_{i-1}, [x, y - 1], f_{i+1}, \ldots, f_{|\mathcal{R}|} \rangle$
          $(C_c, D_c) \leftarrow$ **Closure**(C, D')
          si $(C, D') \ll (C_c, D_c)$ **alors**
            $gauche'[i] \leftarrow faux$
            LoCoM((C_c, D_c), **Propager**$(X, (C_c, D_c))$, $gauche'$)
          LoCoM((C, D), $X \setminus \{(a, [x, y])\}$, $gauche$)

5 Etude empirique

Dans cette section, nous présentons nos principaux résultats expérimentaux. Nous commençons par décrire les jeux de données réels utilisés, ainsi que les questions auxquelles nous voulons répondre. Après l'étude quantitative, nous donnons quelques exemples de motifs trouvés. Pour garantir la reproductibilité, le code source et les données sont librement accessibles [1]. Nous exposons dans cette section qu'un échantillon des études que nous avons réalisées. L'ensemble complet des figures issues de ces expérimentations sont accessibles via le pointeur précédent.

5.1 Jeux de données et objectifs

Nous considérons 4 jeux de données réels bien connus. Le premier [2], **SA-heart**, décrit des individus d'une région d'Afrique du Sud présentant des anomalies cardiaques. Les trois autres jeux de données, issus de la collection UCI Machine Learning [3], décrivent différent domaines d'applications (Abalon, Seismic-bumps, German Credit). Cette étude expérimentale vise à répondre à différentes questions : Comment se comporte LOCOM vis-à-vis des différents paramètres, des caractéristiques des jeux de données et d'une baseline ? des propriétés d'élagage de LOCOM ?

Nous montrons dans le tableau 2 les meilleurs motifs (par rapport à la WRAcc) obtenus par notre approche sur les différents jeux de données.

5.2 Etude quantitative

Comme baseline, nous considérons l'algorithme PAIRMINING (Prado et al. (2013)), dérivé de Calders et al. (2006) afin de gérer les variations positives et négatives. Pour chaque contexte, l'algorithme PAIRMINING recherche des motifs graduels. Les temps d'exécution de LOCOM et PAIRMINING en fonction des paramètres α et β pour le jeu de données SA-heart sont décrits dans la figure 2. Comme attendu, l'algorithme LOCOM est meilleur que PAIRMINING, excepté quand α est suffisamment élevé et que quasiment aucun motif n'est retourné. Plus intéressant, cette baseline ne finit pas dans de nombreuses configurations.

Nous étudions ensuite le comportement de notre algorithme plus en détail. Plus particulièrement, la figure 1 retourne le temps d'exécution, le nombre d'éléments explorés avec ou sans l'exploitation de la borne supérieure $UB_{\textbf{WRAcc}}$ en fonction de k. La distribution des valeurs **WRAcc** des motifs découverts est également affichée sur cette figure. L'optimisation basée sur $UB_{\textbf{WRAcc}}$ permet d'accélérer la découverte des top-k sous-groupes corrélés grâce à un élagage de l'espace de recherche plus efficace. Ce gain atteint même un facteur de 2 sur certains jeux de données.

1. https://www.dropbox.com/sh/arjr651tk5hvqqf/AAB1dJq5Nd9AvVKc6_H5hlnAa?dl=0
2. https://web.stanford.edu/~hastie/ElemStatLearn/datasets/
3. https://archive.ics.uci.edu/ml/index.php

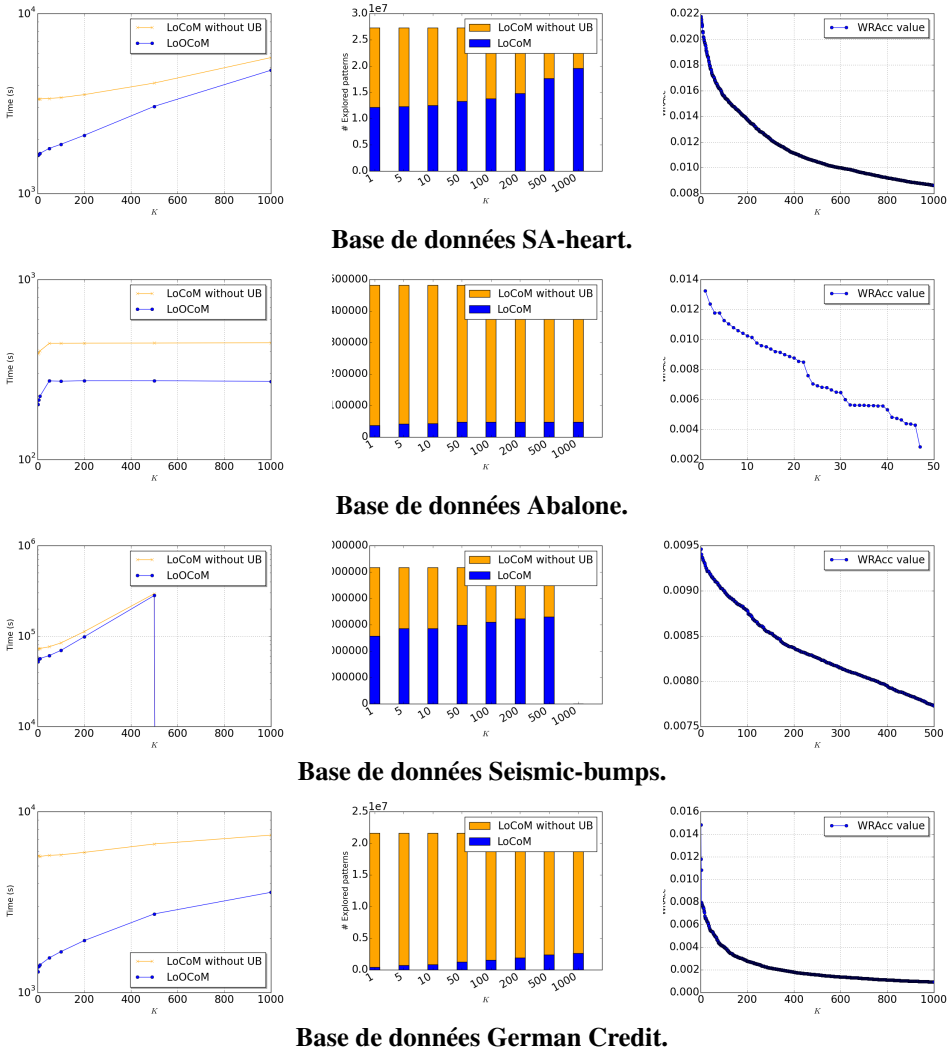
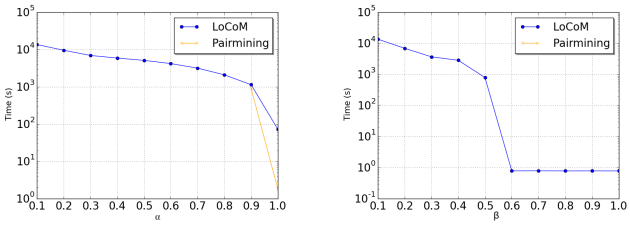


FIG. 1 – *Temps d'exécution (à gauche), nombre de candidats explorés (au centre) et distribution des top-k (à droite) pour* LOCOM *avec et sans optimisation basée sur* $UB_{\mathbf{WRAcc}}$ *selon* k *(*$\alpha = 0.05$ *et* $\beta = 0.4$*).*

5.3 Etude qualitative

Ces motifs sont cohérents par rapport aux connaissances du domaine. Par exemple, pour SA-heart, les maladies coronariennes (chd) sont positivement corrélées avec l'âge et l'obésité. L'âge et la consommation d'alcool sont anti-corrélés pour des valeurs de tensions artérielles (Sdb) élevées. Les motifs découverts sur Abalone mettent en évidence des corrélations entre les différentes mesures de poids sur l'ensemble de données ($\sigma(D) \geq 0,99$).



base de données SA-heart.

FIG. 2 – *Temps d'exécution de* LOCOM *et* PAIRMINING *en fonction de la variation de* α *(gauche) et* β *(droite) (par défaut,* $\alpha = 0.1 = \beta$*).*

C	D	$\tau(C,O)$	$\sigma(D)$	$\tau(C,\sigma(D))$	**WRAcc**
Motifs SA-heart ($\alpha = 0.05$, $\beta = 0.5$ et $\delta = 0.2$).					
Alcohol$^+$, Age$^-$	Sdb = $[121, 194]$	0.512	0.803	0.536	0.015
Age$^+$, Chd$^+$	Sdb = $[101, 138]$	0.591	0.628	0.617	0.010
Obesity$^+$, Chd$^+$	Sdb = $[117, 148]$	0.510	0.665	0.523	0.005
Motifs Abalone ($\alpha = 0.05$, $\beta = 0.5$ et $\delta = 0.7$).					
Shucked weight$^+$, Shell weight$^+$	Diam. =$[0.055, 0.630]$, Height = $[0.01, 0.230]$	0.879	0.996	0.880	0.0005
Shucked weight$^+$, Shell weight$^+$, Rings$^+$	Diam. =$[0.055, 0.545]$, Height = $[0.01, 0.130]$	0.874	0.406	0.876	0.0004
Whole weight$^+$, Shucked weight$^+$	Diam. =$[0.055, 0.650]$, Height = $[0.01, 0.250]$	0.699	0.999	0.699	0.0002
Seismic-bumps patterns ($\alpha = 0.05$, $\beta = 0.5$ et $\delta = 0.8$).					
Number-of-bumps5$^+$, Number-of-bumps6$^+$	shift = coal_setting	0.661	0.643	0.757	0.039
German credit patterns ($\alpha = 0.05$, $\beta = 0.3$ et $\delta = 0.8$).					
Age$^+$, Number-of-credits$^-$	Marital = single, Guarantors = none, Type_of_apartment = own	0.447	0.369	0.503	0.007

TAB. 2 – *Meilleurs motifs par rapport à* ***WRAcc****.*

Ces mesures sont également corrélées avec l'âge de l'abalone (Rings) quand ils sont plutôt petits. Dans les données Seismic-bumps, les nombres de secousses sismiques (Number-of-

bumps) dont l'énergie varie entre $[10^6, 10^7)$ (Number-of-bumps5) et $[10^7, 10^8)$ (Number-of-bumps6) sont corrélés dans des contextes où l'on a de l'extraction de charbon.

Dans les données German credit, l'âge est anti-corrélé avec le nombre de crédits (Number-of-credit) pour les hommes célibataires qui sont propriétaires de leur appartement (Type of apartment = own).

6 Conclusion et perspectives

Dans cet article, nous avons introduit le problème de la découverte des sous-groupes corrélés sur les rangs avec un nombre arbitraire de cibles numériques (supérieur ou égal à 2). Cela permet de mettre en évidence des sous-groupes d'objets – identifiés par des conditions sur des attributs numériques et/ou nominaux – pour lesquels la corrélation de rang entre un sous-groupe d'attributs (numériques ou ordinaux) signés est exceptionnellement supérieure à celle évaluée sur l'ensemble des données. Les motifs de corrélation de rang que nous considérons sont basés sur une généralisation du τ de Kendall qui permet de représenter un sous-ensemble d'attributs numériques qui co-varient d'une manière positive ou négative. Nous avons défini LOCOM, un algorithme de type *Branch-and-Bound* qui exploite certaines propriétés d'élagage basées sur le calcul de bornes supérieures et sur des propriétés de fermeture. Une étude empirique sur plusieurs ensembles de données démontre l'efficacité de LOCOM. Ce travail ouvre de nouvelles perspectives de recherche. Par exemple, d'autres mesures et paradigmes peuvent être étudiées pour évaluer l'intérêt des sous-groupes, en particulier l'intérêt subjectif qui permet de prendre en compte les connaissances a priori de l'utilisateur Bie (2011). Une autre direction intéressante consiste à concevoir des méthodes d'exploration instantanée en abandonnant la complétude de l'algorithme et en échantillonnant directement l'espace des motifs Boley et al. (2011).

Remerciements

Ce travail est partiellement financé par le Labex IMU (Intelligences des Mondes Urbains) dans le cadre du projet RESALI – REseaux et Système ALImentaire Systèmes d'information innovants et exploratoires pour plus de justice alimentaire dans les métropoles (2015).

Références

Bie, T. D. (2011). Maximum entropy models and subjective interestingness : an application to tiles in binary databases. *Data Min. Knowl. Discov. 23*(3), 407–446.

Boley, M., C. Lucchese, D. Paurat, et T. Gärtner (2011). Direct local pattern sampling by efficient two-step random procedures. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pp. 582–590.

Calders, T., B. Goethals, et S. Jaroszewicz (2006). Mining rank-correlated sets of numerical attributes. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 96–105. ACM.

Do, T. D. T., A. Laurent, et A. Termier (2010). PGLCM : efficient parallel mining of closed frequent gradual itemsets. In *ICDM*, pp. 138–147.

Do, T. D. T., A. Termier, A. Laurent, B. Negrevergne, B. Omidvar-Tehrani, et S. Amer-Yahia (2015). Pglcm : efficient parallel mining of closed frequent gradual itemsets. *Knowledge and Information Systems 43*(3), 497–527.

Downar, L. et W. Duivesteijn (2017). Exceptionally monotone models—the rank correlation model class for exceptional model mining. *Knowledge and Information Systems 51*(2), 369–394.

Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. In *PKDD*, pp. 200–211.

Kaytoue, M., S. O. Kuznetsov, et A. Napoli (2011). Revisiting numerical pattern mining with formal concept analysis. In T. Walsh (Ed.), *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pp. 1342–1347. IJCAI/AAAI.

Klösgen, W. (1996). Explora : A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pp. 249–271. AAAI.

Lavrač, N., P. Flach, et B. Zupan (1999). Rule evaluation measures : A unifying view. In S. Džeroski et P. Flach (Eds.), *Inductive Logic Programming*, Berlin, Heidelberg, pp. 174–185. Springer Berlin Heidelberg.

Prado, A., M. Plantevit, C. Robardet, et J.-F. Boulicaut (2013). Mining graph topological patterns : Finding covariations among vertex descriptors. *IEEE Transactions on Knowledge and Data Engineering 25*(9), 2090–2104.

Wille, R. (1982). Restructuring lattice theory : an approach based on hierarchies of concepts. In *Ordered sets*, pp. 445–470. Springer.

Xin, D., H. Cheng, X. Yan, et J. Han (2006). Extracting redundancy-aware top-k patterns. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 444–453. ACM.

Summary

Subgroup discovery (SD) is a mature field at the frontier of data mining and machine learning. It gathers methods designed to find coherent subgroups of a dataset where one or more targets interact in an unusual way. Correlation model classes have already been defined to discover interesting subgroups when dealing with two numerical targets. However, in this supervised setting, the two numerical targets are fixed before the subgroup search. To make unsupervised exploration possible, we propose to search for arbitrary subsets of numerical targets whose correlation is exceptional for an automatically found subgroup. We introduce the problem of rank-correlated subgroup discovery with an arbitrary subset of numerical targets. A rank-correlated subgroup is identified by both conditions on descriptive attributes, whether numeric or nominal, and a pattern on numeric attributes that captures (positive or negative) rank correlations. We define a new branch-and-bound algorithm that exploits some pruning properties. An empirical study on several datasets demonstrates the efficiency and the effectiveness of the algorithm.