Découverte d'un sous-groupe optimal dans des données purement numériques

Alexandre Millot, Rémy Cazabet, Jean-François Boulicaut

Découverte d'un sous-groupe optimal dans des données purement numériques

Alexandre Millot*, Rémy Cazabet**, Jean-François Boulicaut*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
**Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
{prenom.nom}@insa-lyon.fr, {prenom.nom}@univ-lyon1.fr

Résumé. La découverte de sous-groupes dans des données étiquetées consiste à calculer des motifs dans un espace de description des objets pour faire émerger des ensembles d'objets qui ont une répartition particulière du point de vue des étiquettes, par exemple la surreprésentation d'une valeur. Découvrir des sous-groupes intéressants dans des données purement numériques - attributs et étiquette cible - a été peu traité. Généralement, on exploite des discrétisations qui engendrent une perte d'information et des résultats sous-optimaux. Nous traitons le problème du calcul d'un sous-groupe optimal au regard d'une mesure de qualité dans des données purement numériques. Nous exploitons des concepts de fermetures sur des motifs d'intervalles et des techniques d'élagage sophistiquées. Nous validons empiriquement la pertinence de notre algorithme et décrivons succinctement un cas d'application à l'optimisation de la pousse de végétaux en environnement contrôlé.

1 Introduction

La fouille de données numériques est pertinente dans de nombreux contextes applicatifs. On dispose alors de données sur des objets décrits par des valeurs d'attributs numériques. On peut considérer que l'un de ces attributs est une étiquette cible et vouloir mettre en oeuvre soit de l'apprentissage de modèles prédictifs de la valeur de cette étiquette pour de nouveaux objets soit, ce qui va nous intéresser ici, des méthodes de découverte de sous-groupes (Wrobel (1997)). Cette tâche consiste à chercher des sous-ensembles d'objets, des sous-groupes, démontrant des caractéristiques intéressantes d'après une mesure de qualité calculée sur une étiquette cible. La mesure de qualité doit capturer des différences sur la distribution de l'étiquette cible entre le sous-ensemble d'objets considéré et l'ensemble des objets du jeu de données. Un large éventail de méthodes exhaustives (Atzmueller et Puppe (2006) ; Grosskreutz et Paurat (2011)) et heuristiques (Mampaey et al. (2012) ; Bosc et al. (2017)) ont été proposées. Dans la majorité des cas, les approches considèrent des attributs nominaux et une étiquette binaire. Pour ce qui est du traitement d'attributs numériques, quelques travaux (Grosskreutz et Rüping (2009) ; Nguyen et Vreeken (2016)) présentent des méthodes permettant d'éviter une simple discrétisation des attributs. Malgré tout, aucune des méthodes que nous connaissons ne propose de solution exhaustive (et donc la possibilité d'identifier un optimum global de la mesure

de qualité) n'incluant pas de discrétisation sous une forme ou sous une autre. Lorsque l'étiquette cible est numérique, (Lemmerich et al. (2016)) propose des mesures de qualité adaptées et l'algorithme `SD-Map*`. C'est l'algorithme de référence du domaine qui cependant requiert une discrétisation préalable des attributs décrivant les objets. Nous ne connaissons donc pas de méthode qui garantisse la découverte du sous-groupe optimal au regard d'une mesure de qualité choisie dans des données purement numériques et sans aucune discrétisation préalable. Motivons l'intérêt de cette tâche sur l'exemple de l'optimisation des fermes urbaines (e.g., Agripolis, FUL, UrbanLeaf [1]) pour lesquelles des recettes de pousse de végétaux mettent en jeu des attributs numériques (la température, l'hygrométrie, la concentration en CO2, etc) et une étiquette cible numérique (le poids de la récolte, la consommation énergétique, etc). On souhaite fouiller les traces d'exécution des recettes pour découvrir les vecteurs caractéristiques d'une pousse optimisée. Entre les mains d'experts, ces caractéristiques pourront être exploitées pour définir de meilleures recettes. Dans un tel contexte, la découverte garantie d'un sous-ensemble de recettes optimal vis-à-vis de l'étiquette cible est plus pertinente que la découverte, éventuellement heuristique et donc sans garantie, des k meilleurs ensembles.

Nous allons travailler sur les espaces de recherche des motifs d'intervalles introduits dans (Kaytoue et al. (2011)). Notre contribution consiste en un algorithme d'énumération exhaustive de tous les motifs d'intervalles de l'espace de recherche. Notre approche (i) exploite un système de fermeture sur les positifs adapté au contexte numérique pour opérer dans un sous-espace (ii) emploie une nouvelle borne supérieure serrée plus rapide à évaluer et généralisable à plusieurs mesures de qualité (iii) utilise plusieurs techniques avancées d'élagage (*forward checking*, réordonnancement de branches). Le résultat est un algorithme efficace permettant de découvrir un sous-groupe globalement optimal dans des données purement numériques sans discrétisation préalable. La Section 2 formalise le problème traité. Dans la Section 3, nous discutons des travaux existants et de leurs limitations. Nos contributions sont décrites dans la Section 4 avant de les valider empiriquement et de proposer un cas d'application dans la Section 5. La Section 6 conclut brièvement.

2 Formalisation du problème traité

Jeu de données purement numérique. Un jeu de données purement numérique (G, M, T) est composé d'un ensemble d'objets G, d'un ensemble d'attributs numériques M et d'une étiquette cible numérique T. Pour un jeu de données, le domaine des valeurs prises par un attribut $m \in M$ est un ensemble fini ordonné noté D_m. Dans ce contexte, $m(g) = d$ signifie que pour l'attribut m, l'objet g a pour valeur d. Le domaine des valeurs prises par l'étiquette cible T est lui aussi un ensemble fini ordonné noté D_T. $T(g) = v$ signifie que l'étiquette cible T de l'objet g a pour valeur v. La Table 1 représente un jeu composé de deux attributs $M = \{m_1, m_2\}$ et d'une étiquette cible T. Un sous-groupe p est défini par un motif, i.e., sa description ou intention, et l'ensemble des objets du jeu de données où il apparaît, i.e., son extension, notée $ext(p)$. Par exemple, dans la Figure 1, le domaine de valeurs de m_1 est $\{1, 2, 3, 4\}$ et le motif $\langle [2, 4], [1, 3] \rangle$ désigne un sous-groupe dont l'extension est $\{g_3, g_4, g_5, g_6\}$.

1. `http://agripolis.eu/`, `http://www.fermeful.com/`, `https://www.urban-leaf.com/`.

Motif d'intervalles, extension et fermeture. Dans un jeu de données purement numérique (G, M, T), un motif d'intervalles p est un vecteur d'intervalles tel que $p = \langle [b_i, c_i] \rangle_{i \in \{1,...,|M|\}}$ où $b_i, c_i \in D_{mi}$, chaque intervalle correspond à une restriction sur un attribut de M, et $|M|$ est le nombre d'attributs. Un objet $g \in G$ fait partie de l'extension d'un motif d'intervalles $p = \langle [b_i, c_i] \rangle_{i \in \{1,...,|M|\}}$ lorsque $\forall i \in \{1, ..., |M|\}, m_i(g) \in [b_i, c_i]$. Soit p_1 et p_2 deux motifs d'intervalles. $p_1 \subseteq p_2$ signifie que p_2 renferme p_1, i.e., l'hyper-rectangle de p_1 est inclus dans celui de p_2. On dit que p_1 est une spécialisation de p_2. Étant donnés un motif d'intervalles p et son extension $ext(p)$, p est dit clos s'il représente le plus petit ensemble (i.e., le plus petit hyper-rectangle) contenant $ext(p)$. La Figure 1 représente le jeu de la Table 1 dans un espace cartésien ainsi qu'une comparaison entre deux motifs d'intervalles non clos (c_1) et clos (c_2).

	m_1	m_2	T
g_1	1	1	15
g_2	1	2	30
g_3	2	2	60
g_4	3	2	40
g_5	3	3	70
g_6	4	3	85

TAB. 1*: Jeu de données purement numérique.*

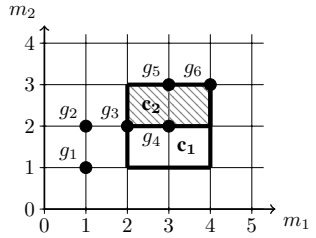

FIG. 1*: Motifs d'intervalles non clos (c_1 = $\langle [2,4], [1,3] \rangle$, non hachuré) et clos ($c_2$ = $\langle [2,4], [2,3] \rangle$, hachuré).*

Mesure de qualité et borne supérieure. Étant donné un jeu de données purement numérique (G, M, T), à chaque motif d'intervalles est associé une valeur numérique permettant d'évaluer sa qualité. Généralement, cette valeur correspond à la différence de distribution de l'étiquette cible entre le jeu de données complet et l'extension du motif d'intervalles considéré. Nous utilisons la famille de mesures basée sur la moyenne introduite dans (Lemmerich et al. (2016)) : Soit p un motif d'intervalles et $ext(p)$ son extension. La qualité de p est calculée via la formule :

$$q^a_{mean}(p) = |ext(p)|^a \times (\mu_{ext(p)} - \mu_{ext(\emptyset)}), a \in [0, 1]$$

où $\mu_{ext(p)}$ est la moyenne de l'étiquette cible pour p, $\mu_{ext(\emptyset)}$ est la moyenne de l'étiquette cible sur l'ensemble du jeu de données, $|ext(p)|$ est la cardinalité de $ext(p)$ et a un paramètre permettant de contrôler le nombre d'objets des sous-groupes retournés.

Étant donné un motif d'intervalles p et une mesure de qualité q, une borne supérieure pour q, notée bs_q, est une fonction permettant de borner la qualité maximale de toutes les spécialisations de p. Formellement, on a $\forall s \subseteq p : q(s) \leq bs_q(p)$. Les bornes supérieures sont utilisées pour élaguer l'espace de recherche : si la borne supérieure d'un motif d'intervalles est inférieure à la qualité minimale requise, il est inutile de considérer les spécialisations du motif.

Sous-groupe optimal. Soit (G, M, T) un jeu de données purement numérique, q une mesure de qualité et P l'ensemble des motifs d'intervalles dans (G, M, T). Un motif d'intervalles p est dit optimal s.s.i $\forall p' \in P : q(p') \leq q(p)$. Notons que plusieurs sous-groupes peuvent avoir la même qualité optimale. Dans cet article, nous retournons le premier trouvé par l'algorithme.

3 État de l'art

Nous ne connaissons pas de solution décrite dans la littérature pour notre problème de découverte d'un sous-groupe optimal dans des données purement numériques mais des problèmes connexes ont été étudiés. Traditionnellement, la recherche de sous-groupes a souvent porté sur des attributs nominaux avec une étiquette cible binaire. Pour traiter des attributs numériques, il faut alors proposer une discrétisation préalable (e.g., Fayyad et Irani (1993)). Les étiquettes cibles numériques peuvent également être discrétisées de façon à se retrouver dans le cas binaire ou nominal (Moreland et Truemper (2009)). Cependant, les discrétisations entraînent des pertes d'informations et il est de fait impossible d'obtenir des garanties sur l'optimalité des sous-groupes obtenus. (Aumann et Lindell (1999)) a introduit la notion de règle d'association quantitative où le conséquent correspond à la moyenne ou à la variance d'un attribut numérique. Une règle est alors définie comme intéressante si sa moyenne dévie significativement de la moyenne de l'ensemble du jeu de données. Par la suite, (Webb (2001)) a proposé une extension de ce type de règle sous le nom de règles d'impact. Ces méthodes ne supportent pas l'énumération exhaustive des sous-groupes sans discrétisation préalable et ne garantissent donc pas la découverte des sous-groupes optimaux. Un problème récurrent pour les algorithmes exhaustifs de fouille de motifs est la taille de l'espace de recherche exponentielle sur le nombre d'attributs. On sait heureusement élaguer dans ces espaces à l'aide de bornes supérieures (Wrobel (1997) ; Grosskreutz et al. (2008)). Les travaux détaillés dans (Lemmerich et al. (2016)) définissent un ensemble de mesures de qualité et de bornes supérieures pour l'énumération exhaustive avec des étiquettes cibles numériques. On trouve de nombreuses propositions (Grosskreutz et Rüping (2009) ; Mampaey et al. (2012) ; Nguyen et Vreeken (2016)) pour traiter des attributs numériques mais ces méthodes se basent toujours sur une forme de discrétisation. L'algorithme `MinIntChange` (Kaytoue et al. (2011)) se base sur des constructions de l'Analyse Formelle de Concepts (Ganter et Wille (1998)) pour la recherche exhaustive de motifs fréquents - et non de sous-groupes - dans des données numériques et sans discrétisation préalable. L'utilisation de systèmes de fermeture et de classes d'équivalence (Soulet et al. (2004) ; Garriga et al. (2008) ; Grosskreutz et Paurat (2011)) est une solution courante pour réduire la taille de l'espace de recherche des sous-groupes. (Belfodil et al. (2018)) a proposé un algorithme étendant des principes de `MinIntChange` à la découverte de sous-groupes en présence d'étiquettes cibles binaires. Nous exploitons les notions de fermeture, de borne supérieure et les principes d'énumération des motifs d'intervalle de `MinIntChange` pour découvrir sans discrétisation les sous-groupes optimaux dans des données purement numériques.

4 Découverte d'un sous-groupe optimal

4.1 Fermeture sur les positifs

La fermeture sur les motifs d'intervalles de (Kaytoue et al. (2011)) a déjà été étendue aux fermetures sur les positifs (Belfodil et al. (2018)).

Définition 1. Soit $p \in P$ un motif d'intervalles, $p' \subseteq p$, $ext(p)$ l'extension de p, T une étiquette cible binaire. Un objet est dit positif si la valeur de son étiquette cible correspond à

la classe que l'on cherche à discriminer et négatif dans le cas inverse. Soit $ext(p)^+$ le sous-ensemble d'objets de $ext(p)$ pour lesquels l'étiquette T est positive. p' est dit clos sur les positifs s'il représente le plus petit ensemble contenant $ext(p)^+$. Soit q une mesure de qualité, on a $q(p) \leq q(p')$. L'idée est que, pour tout sous-groupe $p \in P$, si l'on retire à l'extension de p tous les objets négatifs qui ne sont pas dans l'extension de p', alors la qualité du sous-groupe ne peut diminuer. Notons que les clos sur les positifs sont un sous-ensemble des clos.

La notion de clos sur les positifs pour des étiquettes cibles binaires est extensible aux étiquettes cibles numériques pour un ensemble de mesures de qualité comme q^a_{mean}. Nous transformons l'étiquette numérique en une étiquette binaire : les objets dont la moyenne est strictement supérieure à la moyenne du jeu de données sont définies comme positives et les autres reçoivent une étiquette négative. Notons que la mesure de qualité est calculée sur l'étiquette numérique. La binarisation n'est utilisée que pour améliorer l'élagage dans l'espace de recherche et n'entraîne pas de perte d'information vis à vis des motifs (i.e., on a la garantie de trouver un motif optimal sans discrétisation). La Table 2 représente les données de la Table 1 avec l'étiquette T (moyenne = 50) transformée en étiquette binaire T_b. La Figure 2 représente les données de la Table 2 dans un espace cartésien et une comparaison entre deux motifs d'intervalles clos (c_1) et clos sur les positifs (c_2).

Pour le cas d'un sous-groupe avec mesure de qualité positive, on montre que la qualité du sous-groupe est toujours supérieure ou égale si l'on supprime tous les objets négatifs ne faisant pas partie de la fermeture sur les positifs. Le cas d'un sous-groupe avec une qualité négative est plus complexe à traiter car la fermeture sur les positifs peut entraîner une diminution de la qualité du sous-groupe. On montre que les objets ne faisant pas partie de la fermeture sur les positifs ne peuvent jamais appartenir à la meilleure spécialisation du sous-groupe.

Théorème 1. Soit p un motif d'intervalles, q^a_{mean} un ensemble de mesures de qualité, p^+ la fermeture sur les positifs de p tel que $p^+ \subseteq p$, et $q^a_{mean}(p) \geq 0$, on a $q^a_{mean}(p^+) \geq q^a_{mean}(p),\ a \in [0,1]$.

Preuve. Soit $ext(p)$ l'extension de p, $ext(p)^+$ l'extension de p^+, $ext(p)^- = ext(p) \setminus ext(p)^+$ l'ensemble des objets négatifs de p n'appartenant pas à $ext(p)^+$, et $T(i)$ la valeur de l'étiquette cible pour l'objet i. Pour raccourcir, nous définissons $e = ext(p)$ et $\theta = ext(\emptyset)$.
Nous cherchons à prouver que :

$$|e^+|^a \times (\mu_{e^+} - \mu_\theta) \geq |e|^a \times (\mu_e - \mu_\theta) \tag{1}$$

	m_1	m_2	T	T_b
g_1	1	1	15	−
g_2	1	2	30	−
g_3	2	2	60	+
g_4	3	2	40	−
g_5	3	3	70	+
g_6	4	3	85	+

TAB. 2*: Jeu de données purement numérique avec étiquette binaire (T_b).*

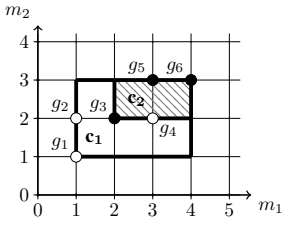

FIG. 2*: Motifs d'intervalles clos (c_1 = $\langle [1,4],[1,3] \rangle$, non hachuré) et clos sur les positifs (c_2 = $\langle [2,4],[2,3] \rangle$, hachuré).*

Ce qui peut être transformé en :

$$|e^+|^a \times \frac{\sum_{i \in e^+}(T(i) - \mu_\theta)}{|e^+|} \geq |e|^a \times \frac{\sum_{i \in e}(T(i) - \mu_\theta)}{|e|} \tag{2}$$

$$|e^+|^a \times \frac{\sum_{i \in e^+}(T(i) - \mu_\theta)}{|e^+|} \geq (|e^+| + |e^-|)^a \times \frac{\sum_{i \in e^+}(T(i) - \mu_\theta) + \sum_{i \in e^-}(T(i) - \mu_\theta)}{|e^+| + |e^-|} \tag{3}$$

Par construction, on sait que $\sum_{i \in e^+}(T(i) - \mu_\theta) \geq 0 \geq \sum_{i \in e^-}(T(i) - \mu_\theta)$. Le reste de la preuve suit de la même façon que (Lemmerich et al. (2016)). On en déduit que pour tout sous-groupe satisfaisant $q^a_{mean}(p) \geq 0$, la fermeture sur les positifs résulte toujours en un sous-groupe de qualité supérieure ou égale.

Théorème 2. Soit p un motif d'intervalles, $ext(p)$ son extension, p^+ la fermeture sur les positifs de p tel que $p^+ \subseteq p$ et $ext(p)^+$ son extension avec $|ext(p)^+| > 0$. Soit $ext(p)^- = ext(p) \setminus ext(p)^+$ l'ensemble des objets négatifs de p n'appartenant pas à $ext(p)^+$ et q^a_{mean} un ensemble de mesures de qualité avec $q^a_{mean}(p) < 0$: Aucun objet appartenant à $ext(p)^-$ ne fait partie de la meilleure spécialisation de p.
$Preuve$. On part du principe qu'il existe un objet de $ext(p)^-$ noté i^- appartenant à la meilleure spécialisation de p, notée p_{top}. Par construction, on a $q^a_{mean}(p_{top}) > 0$ (car $|ext(p)^+| > 0$). Soit p^+_{top} la fermeture de p_{top}. Par construction, on sait que i^- ne fait pas partie de l'extension de p^+_{top}. Or, d'après le Théorème 1, on sait que $q^a_{mean}(p^+_{top}) \geq q^a_{mean}(p_{top})$. On en déduit que i^- n'appartient pas à la meilleure spécialisation de p.

4.2 Borne supérieure serrée

Nous définissons maintenant une nouvelle borne supérieure serrée pour la famille de mesures q^a_{mean}. Une borne supérieure est dite serrée, si, pour tout sous-groupe d'un jeu de données, il existe un sous-ensemble d'objets du sous-groupe dont la qualité est égale à la valeur de la borne supérieure du sous-groupe. Notons qu'il n'est pas nécessaire que le sous-ensemble corresponde à un sous-groupe. Il est possible de calculer une borne supérieure serrée pour les mesures q^a_{mean} en ne passant qu'une seule fois sur chaque objet d'un sous-groupe.

Définition 2. Soit p un motif d'intervalles et $ext(p)$ son extension. Soit $S_i \subseteq ext(p)$ le sous-ensemble d'objets contenant les i objets avec la valeur d'étiquette la plus élevée. Alors, d'après (Lemmerich et al. (2016)), une borne supérieure serrée pour q^a_{mean} est donnée par :

$$bss^a_{mean}(p) = max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)|})), a \in [0, 1]$$

On peut calculer une meilleure borne serrée en se limitant à l'examen des objets positifs.

Théorème 3. Soit p un motif d'intervalles et $ext(p)^+$ l'ensemble des objets de l'extension de p dont la valeur d'étiquette est supérieure à la moyenne du jeu de données. Soit $S_i \subseteq ext(p)^+$ le sous-ensemble d'objets contenant les i objets avec la valeur d'étiquette la plus élevée. Une borne supérieure serrée pour q^a_{mean} est donnée par :

$$\overline{bss}^a_{mean}(p) = max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)^+|})), a \in [0, 1]$$

$Preuve$. Nous devons prouver que :

$$\overline{bss}^a_{mean}(p) \geq bss^a_{mean}(p), a \in [0,1]$$

Autrement dit, il suffit de prouver que : $\forall S_i \subseteq ext(p), q^a_{mean}(S_i^+) \geq q^a_{mean}(S_i)$ avec S_i^+ le sous-ensemble d'objets positifs de S_i. Dans (Lemmerich et al. (2016)), il est prouvé qu'aucun objet négatif ne fait partie du meilleur sous-ensemble d'objets d'un sous-groupe pour la famille de mesures q^a_{mean}. Il suit logiquement que pour tout sous-ensemble S_i, la suppression des objets négatifs ne peut faire diminuer la qualité du sous-ensemble. On a donc :

$$\forall S_i \subseteq ext(p), q^a_{mean}(S_i^+) \geq q^a_{mean}(S_i)$$

On en déduit que :

$$max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)^+|})) \geq max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)|})), a \in [0,1]$$

Donc $\overline{bss}^a_{mean}(p)$ est une borne supérieure serrée pour q^a_{mean}.

4.3 Algorithme

Nous introduisons `OSMIND` pour une recherche en profondeur d'abord d'un sous-groupe optimal. La découverte s'effectue à travers le parcours des motifs d'intervalles clos sur les positifs couplé à l'exploitation de bornes supérieures serrées et de méthodes poussées d'élagage de l'espace de recherche. Le pseudo-code est dans l'Algorithme 1.

Pour garantir la découverte d'un sous-groupe optimal, nous adoptons le concept de changement minimal sur un motif d'intervalles de `MinIntChange` pour assurer une énumération exhaustive. Un changement minimal à droite correspond à remplacer la partie droite d'un intervalle par la valeur inférieure la plus proche de la valeur courante dans le domaine de valeur de l'attribut correspondant. Suivant la même logique, un changement minimal à gauche consiste à remplacer la partie de gauche d'un intervalle par la valeur supérieure la plus proche. La recherche commence avec le motif d'intervalles minimal couvrant la totalité des objets. Le principe est d'appliquer des changements minimaux consécutifs à gauche ou à droite jusqu'à obtenir un intervalle dont les bornes de gauche et de droite ont la même valeur pour chaque intervalle du motif d'intervalles minimal. Dans ce cas, l'algorithme remonte l'arbre des motifs générés jusqu'à trouver un motif sur lequel il est possible d'effectuer un changement minimal.

Nous utilisons la notion de clos sur les positifs adaptée aux étiquettes numériques pour réduire significativement le nombre de motifs d'intervalles candidats. Après chaque changement minimal effectué (Ligne 4), au lieu d'évaluer le nouveau motif d'intervalles résultant, nous calculons et évaluons le motif d'intervalles clos sur les positifs correspondant (Ligne 5).

En effectuant un parcours exhaustif de tous les motifs d'intervalles clos sur les positifs, un motif d'intervalles peut être généré plusieurs fois. Pour éviter cette redondance et assurer l'unicité de la génération du motif, une solution couramment utilisée est l'utilisation d'un test de canonicité. Dans le cas des motifs d'intervalles, le test de canonicité permet de vérifier que la fermeture n'a pas entraîné de changement sur l'un des intervalles précédant l'intervalle sur lequel le changement minimal a été effectué (Ligne 6). Une autre forme de redondance est possible. L'application successive de changements minimaux à gauche ou à droite sur un intervalle entraîne de multiples générations d'un même motif d'intervalles. Une solution est

Algorithme 1 Algorithme OSMIND

```
1: fonction OSMIND( )
2:     Initialise(motif_intervalles_minimal, motif_optimal)
3:     RECURSION(motif_intervalles_minimal, 0)
4:     retourne motif_optimal
5: fin fonction
1: procédure RECURSION(motif, attribut)
2:     pour i de attribut à nb_attributs − 1 faire
3:         pour elem dans {droite, gauche} faire
4:             motif ← changementMinimal(motif, i, elem)
5:             motif_clos ← calculerClosSurPositifs(motif)
6:             si isCanonique(motif_clos) alors
7:                 si borneSerree(motif_clos) > qualite(motif_optimal) alors
8:                 stocker(motif_clos, i) fin si
9:                 si qualite(motif_clos) > qualite(motif_optimal) alors
10:                motif_optimal ← motif_clos fin si
11:            fin si
12:        fin pour
13:    fin pour
14:    pour chaque element stocke ordonne par valeur de borne serree faire
15:        si borneSerree(element.motif) > qualite(motif_optimal) alors
16:        RECURSION(element.motif, element.attribut) fin si
17:    fin pour
18: fin procédure
```

l'application d'une contrainte sur les changements minimaux. Après un changement minimal à droite, un changement minimal à droite ou à gauche peut être appliqué. En revanche, un changement minimal à gauche doit toujours être suivi par un changement minimal à gauche.

Nous devons bien sûr veiller aux possibilités d'élagage de l'espace de recherche. C'est ce que va nous apporter la borne supérieure serrée sur la qualité des spécialisations d'un motif d'intervalles clos sur les positifs. Pour chaque sous-groupe, une borne supérieure est calculée (Ligne 7), et, si celle-ci est inférieure à la qualité du meilleur sous-groupe, l'espace est élagué en ignorant l'ensemble des spécialisations du motif d'intervalles. La seconde technique que nous mettons en oeuvre est le couplage d'un $forward\ checking$ et d'un réordonnancement des branches. Pour un motif d'intervalles donné, l'ensemble de ses spécialisations directes (application d'un changement minimal à gauche ou à droite sur chaque intervalle du motif) sont développées - $forward\ checking$ - et celles dont la borne supérieure est meilleure que le meilleur sous-groupe sont stockées (Ligne 8). Un réordonnancement des branches par ordre décroissant de la valeur de la borne supérieure stockée est ensuite effectué (Ligne 14). Réordonner les branches par ordre décroissant de la valeur de la borne supérieure permet d'explorer d'abord les parties les plus prometteuses de l'espace de recherche. Cela permet également un meilleur élagage en élevant plus rapidement la qualité minimale requise.

Jeu de données	Attr	Obj	|Pl|
Bolt	8	40	8.7×10^9
Basketball	4	96	2.3×10^{11}
Airport	4	135	7.1×10^{15}
Body Temp	2	130	1.8×10^3
Pollution	15	60	1.7×10^{42}
RecettesA	9	100	5.1×10^{18}
RecettesB	9	1000	5.1×10^{18}

TAB. 3*: Jeux de données et leurs caractéristiques : nombre d'attributs, nombre d'objets et taille de l'espace de recherche.*

R	T^{P1}	CO_2^{P1}	T^{P2}	CO_2^{P2}	W
r_1	18	800	24	1000	5
r_2	22	1000	27	950	6
r_3	27	1200	28	650	7
r_4	19	600	17	800	3
r_5	24	500	23	450	9
r_6	16	750	19	1300	2
r_7	30	1100	25	900	8

TAB. 4*: Recettes de pousse découpées en 2 phases (P1 et P2), 2 attributs (température et CO2), et une étiquette cible (poids).*

5 Expérimentations

Nous avons d'abord sélectionné 7 jeux de données purement numériques décrits dans la Table 3. Le code source et les jeux de données utilisés pour nos expérimentations sont disponibles à l'url `https://bit.ly/37Ag9UK`. L'implémentation de `SD-Map*` est disponible dans le système VIKAMINE [2]. Les 5 premiers jeux de données (Bolt, Basketball, Airport, Body Temp et Pollution) proviennent du répertoire Bilkent [3]. Les 2 derniers (RecettesA et RecettesB) correspondent à des simulations de pousse de végétaux que nous avons nous-même générées au moyen de l'environnement Python Crop Simulation Environment PCSE [4]. Chaque simulation de pousse est représentée sous la forme d'un ensemble d'attributs numériques - les conditions de pousse (e.g., température, irradiation solaire, etc) - et d'une étiquette cible numérique - le poids des végétaux obtenus. La pousse d'un végétal est découpée en plusieurs périodes de temps de taille égale appelées *phases*. La Table 4 présente des exemples simplifiés de cultures calculées au moyen du simulateur.

Les gains de performance apportés par nos contributions sont résumés dans la Table 5. Les performances du système de fermeture sur les positifs sont comparées avec celles d'un système de fermeture simple (Section 2). Pour chaque jeu de données, nous effectuons une comparaison du nombre de sous-groupes considérés avant l'obtention du motif optimal pour la mesure de qualité q^a_{mean} avec a = 0,5 et a = 1. Dans tous les cas étudiés, l'approche avec fermeture sur les positifs est substantiellement plus efficace que la seconde approche. En effet, notre méthode permet en moyenne de diviser le nombre de sous-groupes évalués par un facteur 20. Nous étudions ensuite le potentiel gain de performance - en termes de temps d'exécution (en secondes) - apporté par la nouvelle borne serrée. Nous comparons cette borne à la borne serrée sur laquelle elle est basée pour l'ensemble des jeux de données et pour les mêmes mesures que précédemment. Notre nouvelle borne est plus efficace dans tous les cas étudiés et peut conduire à une réduction de près de 30% du temps d'exécution global.

2. `http://www.vikamine.org/`
3. `http://funapp.cs.bilkent.edu.tr/DataSets/`
4. `https://pcse.readthedocs.io/en/stable/index.html`

Jeu	a	FSLP	FN	Gain (÷)	SA	SB	Gain (%)
Bolt	0.5	25	118	4.7	0.0062	0.0078	20.5
	1	16	299	19	0.0042	0.0055	23.6
Basketball	0.5	143037	3014506	21	80.5	104	22.6
	1	42548	1121798	26	30.5	39.3	22.4
Airport	0.5	387	12042	35	0.17	0.19	10.5
	1	57	10055	176	0.033	0.037	10.8
Body Temp	0.5	795	1199	1.5	0.53	0.73	27.4
	1	570	865	1.5	0.47	0.53	11.3
Pollution	0.5	100776	-	-	23.9	25	4.4
	1	1289	41662411	32321	0.376	0.408	7.8
RecettesA	0.5	18258	430105	24	8.25	9,84	16.1
	1	1147	24431	21	0.72	0,82	12.2
RecettesB	0.5	324116	854873	2.6	1666	2223	25
	1	5261	17848	3.4	45.8	64,3	28.8

TAB. 5*: Comparaison des approches : Fermeture sur les positifs (FSLP) vs Fermeture normale (FN) et Serrée Améliorée (SA) vs Serrée Base (SB). "-" signifie temps d'exécution >72h.*

Observons maintenant les performances de OSMIND par rapport à l'algorithme de référence SD−Map⋆. Nous décidons de comparer la qualité du meilleur sous-groupe retourné par chacune des deux méthodes sur les 5 premiers jeux de données de la Table 3 pour la mesure de qualité q^a_{mean} avec a = 0,5. Pour SD−Map⋆, une discrétisation préalable des attributs numériques est nécessaire. Dans le but d'obtenir des résultats comparables et impartiaux, nous testons plusieurs techniques de discrétisation avec différents nombres d'intervalles (2, 3, 5, 10, 15 et 20) pour SD−Map⋆, et nous retenons uniquement la meilleure solution pour la comparaison avec OSMIND. Les techniques de discrétisation utilisées sont *Equal-Width*, *Equal-Frequency* et *K-Means*. Les résultats sont présentés dans la Figure 3. Notre algorithme permet d'obtenir des sous-groupes de qualité significativement supérieure pour chaque jeu de données testé, et ce peu importe la discrétisation utilisée par SD−Map⋆. Nous en déduisons que la perte d'information inhérente à la discrétisation des attributs est responsable des moins bons résultats obtenus avec SD−Map⋆.

Comparons maintenant OSMIND avec SD−Map⋆ sur des données de pousse de végétaux en milieu contrôlé. L'environnement PCSE nous permet générer 1000 cultures aléatoires. Nous sélectionnons ensuite 10, 50, 100, 200, 500 et 1000 cultures dans ce jeu de données et nous enregistrons la qualité du meilleur sous-groupe obtenu pour la mesure de qualité q^a_{mean} avec a = 0,5. Pour ce qui est de SD−Map⋆, nous retenons à nouveau la discrétisation produisant le meilleur sous-groupe. La Figure 4 permet d'observer la qualité relative du meilleur sous-groupe retourné par chaque algorithme en fonction de la taille du jeu de données. Pour des jeux de petite taille, SD−Map⋆ produit un sous-groupe globalement optimal malgré la discrétisation appliquée. En revanche, dès que l'on s'intéresse à des jeux de données avec un plus grand nombre d'objets, SD−Map⋆ est constamment 10 à 25% moins performant qu'OSMIND. Intéressons nous à la description du sous-groupe optimal obtenu par OSMIND et SD−Map⋆. Une présentation de ces descriptions pour le jeu de données RecettesA est dans la Table 6. Outre la qualité supérieure du sous-groupe retourné par OSMIND, la description du sous-groupe optimal permet également d'extraire des informations absentes de la description obtenue avec SD−Map⋆. En effet, là où SD−Map⋆ propose uniquement une restriction forte sur l'attribut

$Irrad^{P2}$, OSMIND permet l'obtention d'informations exploitables sur 5 des 9 attributs considérés. Cela indique une supériorité qualitative de notre approche.

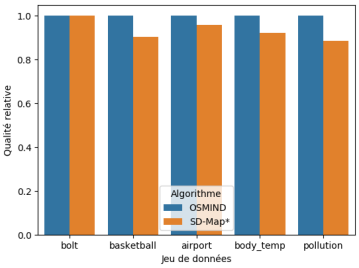

FIG. 3: *Comparaison de la qualité du meilleur sous-groupe.*

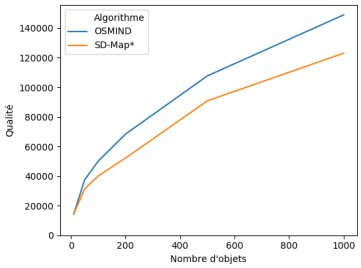

FIG. 4: *Qualité du meilleur sous-groupe en fonction du nombre d'objets.*

Sous-groupe	Pluie^{P1}	Irrad^{P1}	Vent^{P1}	Pluie^{P2}	Irrad^{P2}	Vent^{P2}	Pluie^{P3}	Irrad^{P3}	Vent^{P3}	Q	S
JDD	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	0	100
OSMIND	-	[4428,23285]	[0,27]	[8,40]	[16428,25000]	-	[2,40]	-	-	50147	26
SD-Map*	-	-	-	-	[19000,25000]	-	-	-	-	40069	31

TAB. 6: *Comparaison des descriptions du jeu de données complet (JDD), du sous-groupe optimal retourné par* OSMIND*, et de celui retourné par* SD−Map⋆*. "-" signifie aucune restriction sur l'attribut par rapport au JDD, Q et S respectivement la qualité et la taille du sous-groupe.*

Nous voulons travailler à l'optimisation de recettes dans des fermes urbaines mais nous ne disposons pas encore de données réelles susceptibles d'être traitées. Nous pouvons travailler sur ce cas d'application au moyen du simulateur déjà introduit. En effet, dans une ferme urbaine, les végétaux sont développés en milieu contrôlé (e.g., température, CO_2, etc). Une recette de pousse est l'ensemble des conditions de développement d'un végétal tout au long de sa croissance. À priori, en l'absence de panne, les consignes fixées par la recette sont respectées et l'un des objectifs d'optimisation peut concerner la masse globale de végétaux récoltés à l'issue de l'exécution de la recette. La Table 4 présente des exemples de recettes de pousse en milieu contrôlé. Nous pouvons simuler l'exécution de recettes au moyen de l'environnement de simulation PCSE en fixant les caractéristiques (e.g., du climat) sur les différentes phases spécifiées. Nous avons utilisé le simulateur pour générer 30 recettes avec des conditions de pousses aléatoires. Nous nous intéressons à 3 variables qui fixent les quantités d'irradiation solaire, de vent et de pluie. La pousse des végétaux est ici découpée en 3 phases de taille égale. Nous pouvons d'abord vérifier que OSMIND permet de découvrir un sous-groupe maximisant le poids moyen des végétaux. Nous voulons ensuite confirmer l'interprétabilité et l'exploitabilité des résultats retournés. La Table 7 fournit une comparaison entre le motif d'intervalles du jeu de départ et le motif d'intervalles du sous-groupe optimal obtenu avec OSMIND. Ces résultats confirment la capacité d'OSMIND à découvrir un sous-groupe de recettes avec un poids moyen optimisée (17819 vs 7256). On peut se servir de la description de ce sous-groupe optimal comme une nouvelle recette permettant un meilleur poids récolté. Le motif d'intervalles optimal retourné est facilement interprétable et permet l'extraction de connaissances non triviales. Par exemple,

Sous-groupe	Pluie^{P1}	Irrad^{P1}	Vent^{P1}	Pluie^{P2}	Irrad^{P2}	Vent^{P2}	Pluie^{P3}	Irrad^{P3}	Vent^{P3}	P
JDD	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	7256
OS	[0,40]	[2714,23285]	[0,21]	[8,37]	[16428,25000]	[0,23]	[2,40]	[6142,25000]	[0,27]	17819

TAB. 7*: Résultats d'*OSMIND*. Motifs d'intervalles du jeu de données complet (JDD) et du sous-groupe optimal trouvé (OS), et poids moyen (P) des recettes de chaque sous-groupe.*

pendant la première phase de pousse, la quantité d'irradiation solaire (Irrad^{P1}) à laquelle sont soumis les végétaux ne semble pas avoir d'impact sur l'optimisation du poids récolté. On peut déduire cela à partir de la faible restriction appliquée sur l'intervalle de valeurs pris par Irrad^{P1}. La connaissance du domaine confirme : la capacité d'absorption de lumière des végétaux est limitée pendant la première phase de développement et nous pourrions donc réduire le coût de pousse à rendement égal en limitant la quantité de lumière utilisée en début de cycle.

6 Conclusion

Nous étudions la découverte d'un sous-groupe optimal vis-à-vis d'une mesure de qualité dans des données purement numériques. Nous avons motivé les raisons pour lesquelles les approches existantes produisent généralement des résultats sous-optimaux du fait des discrétisations réalisées. L'algorithme `OSMIND` garantit la découverte d'un sous-groupe optimal. La validation empirique effectuée et l'application à l'optimisation de recettes pour la pousse de végétaux nous ont permis de confirmer la pertinence et l'exploitabilité d'`OSMIND`. Nos futurs travaux pourront porter sur l'adaptation et l'application d'`OSMIND` à des données de grande dimension et le développement du cas d'application à l'optimisation de recettes pour des fermes urbaines. De plus, notre approche permettant de trouver un sous-groupe optimal, il serait intéressant d'explorer son couplage avec des techniques de couverture séquentielle pour la découverte des ensembles de sous-groupes optimaux non-redondants.

Remerciements Cette recherche est partiellement financée par le FUI DUF 4.0 (2017-2021).

Références

Atzmueller, M. et F. Puppe (2006). SD-Map – a fast algorithm for exhaustive subgroup discovery. In *Proceedings PKDD*, pp. 6–17.

Aumann, Y. et Y. Lindell (1999). A statistical theory for quantitative association rules. In *Proceedings ACM SIGKDD*, pp. 261–270.

Belfodil, A., A. Belfodil, et M. Kaytoue (2018). Anytime subgroup discovery in numerical domains with guarantees. In *Proceedings ECML/PKDD*, pp. 500–516.

Bosc, G., J.-F. Boulicaut, C. Raïssi, et M. Kaytoue (2017). Anytime discovery of a diverse set of patterns with monte carlo tree search. *Data Min. Knowl. Discov. 32*, 604–650.

Fayyad, U. M. et K. B. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings IJCAI*, pp. 1022–1029.

Ganter, B. et R. Wille (1998). *Wille, R. : Formal Concept Analysis : Mathematical Foundations.* Springer.

Garriga, G. C., P. Kralj, et N. Lavrač (2008). Closed sets for labeled data. *J. Mach. Learn. Res. 9*, 559–580.

Grosskreutz, H. et D. Paurat (2011). Fast and memory-efficient discovery of the top-k relevant subgroups in a reduced candidate space. In *Proceedings ECML/PKDD*, pp. 533–548.

Grosskreutz, H. et S. Rüping (2009). On subgroup discovery in numerical domains. *Data Min. Knowl. Discov. 19*(2), 210–226.

Grosskreutz, H., S. Rüping, et S. Wrobel (2008). Tight optimistic estimates for fast subgroup discovery. In *Proceedings ECML/PKDD*, pp. 440–456.

Kaytoue, M., S. O. Kuznetsov, et A. Napoli (2011). Revisiting numerical pattern mining with formal concept analysis. In *Proceedings IJCAI*, pp. 1342–1347.

Lemmerich, F., M. Atzmueller, et F. Puppe (2016). Fast exhaustive subgroup discovery with numerical target concepts. *Data Min. Knowl. Discov. 30*(3), 711–762.

Mampaey, M., S. Nijssen, A. Feelders, et A. Knobbe (2012). Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data. In *Proceedings IEEE ICDM*, pp. 499–508.

Moreland, K. et K. Truemper (2009). Discretization of target attributes for subgroup discovery. In *Proceedings MLDM*, pp. 44–52.

Nguyen, H. V. et J. Vreeken (2016). Flexibly mining better subgroups. In *Proceedings SIAM SDM*, pp. 585–593.

Soulet, A., B. Crémilleux, et F. Rioult (2004). Condensed representation of emerging patterns. In *Proceedings PAKDD*, pp. 127–132.

Webb, G. I. (2001). Discovering associations with numeric variables. In *Proceedings ACM SIGKDD*, pp. 383–388.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings PKDD*, pp. 78–87.

Summary

Subgroup discovery in labeled data is the task of discovering patterns in the description space of objects to find subsets of objects whose labels show an interesting distribution, for example the disproportionate representation of a label value. Discovering interesting subgroups in purely numerical data - attributes and target label - has received little attention so far. Usually, one uses discretization methods that lead to a loss of information and suboptimal results. We consider the discovery of an optimal subgroup according to an interestingness measure in purely numerical data. We leverage concepts of closures on interval patterns and advanced pruning techniques. The relevance of our algorithm is studied empirically and we briefly describe an application scenario to the optimization of plant growth in a controlled environment.