Une approche déclarative pour le clustering explicable sous contraintes

Mathieu Guilbert*, Christel Vrain*,Thi-Bich-Hanh Dao*

*Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022, Orléans, France
prenom.nom@univ-orleans.fr

Résumé. Le clustering est souvent considéré comme une tâche non supervisée, exploratoire, aidant un expert à comprendre la structure de ses données. Des contraintes fondées sur des connaissances expertes peuvent être introduites afin d'aligner les résultats avec ses attentes. Cependant leur acquisition reste complexe, rendant l'explication du clustering essentielle pour ajuster les paramètres et découvrir de nouvelles informations. Nous abordons le problème du clustering explicable en considérant deux espaces de représentation : l'un pour le clustering et l'autre pour les explications. Notre méthode ECS (Explainability-driven Cluster Selection) vise à produire un clustering de qualité, tout en étant explicable par des motifs couvrant la majorité des instances d'un cluster et les discriminant des autres. Elle s'appuie sur le clustering ensemble et un nouveau modèle de programmation par contraintes (PPC) pour la sélection des clusters et des explications.

1 Introduction

Le clustering est une tâche non-supervisée d'apprentissage automatique dont le but est d'organiser les données en différents groupes (ou clusters) regroupant des observations similaires entre elles et différentes des observations des autres clusters. Il est souvent vu comme une tâche exploratoire, permettant aux experts de mieux comprendre la structure sous-jacente de leurs données. Les résultats obtenus par un algorithme de clustering dépendent de nombreux facteurs. Des connaissances expertes, exprimées sous forme de contraintes, peuvent être introduites dans l'objectif de rapprocher les résultats des attentes des experts (voir par exemple (Dao et Vrain, 2024)). Cependant, l'obtention de ces contraintes est difficile, rendant ainsi essentiel d'expliquer les résultats d'un clustering non seulement pour guider le paramétrage des algorithmes, mais aussi afin d'assister l'expert pour proposer de nouvelles connaissances.

Nous abordons la problématique du clustering explicable dans un cadre où les données sont décrites dans deux espaces de représentation (identiques ou non) : un espace de propriétés pour le clustering et un espace booléen pour les explications. L'utilisation de deux espaces nous permet de proposer une approche générique englobant deux situations : lorsque le même espace est utilisé pour le clustering et les explications, nécessitant dans ce cas de discrétiser les attributs numériques, ou lorsqu'un autre ensemble de descripteurs semble pertinent pour expliquer, permettant de mettre en évidence des relations entre ces espaces. En guise d'illustration, considérons une application médicale où des patients souffrant d'une maladie se sont vus attribuer

un score représentant la gravité de leur état et sont décrits par différentes propriétés cliniques (âge, stress, ...). Dans ce contexte, il serait utile de classer les patients en fonction de leur état de santé et d'interpréter ces groupes à l'aide de propriétés cliniques booléennes, mettant ainsi en évidence des relations entre la sévérité de la maladie et ces propriétés cliniques.

La plupart des approches de clustering, comme K-Means, reposent sur une distance entre observations. Cependant, il existe une famille moins connue d'approches appelée clustering conceptuel introduite par (Fisher, 1987; Michalski et Stepp, 1983) et fondée sur la notion de concept. Elle présente un grand intérêt pour le clustering explicable car elle repose sur l'idée qu'un cluster doit être un concept, c'est-à-dire un ensemble d'observations couplé à un ensemble de propriétés exclusivement satisfaites par les observations du cluster. Néanmoins, cette notion de concept est souvent trop forte pour des applications concrètes et doit être relâchée, comme montrée dans (Dao et al., 2018). Notre proposition fondé sur deux espaces de représentation s'inscrit dans le cadre du clustering descriptif et permet de faire le pont entre le clustering classique fondé sur une distance et le clustering conceptuel, bénéficiant ainsi des atouts des deux approches. Dans notre proposition, un cluster est un ensemble d'observations similaires dans l'espace de clustering, expliqué par un ensemble de motifs (conjonction de descripteurs booléens) dans l'espace d'explication caractéristiques de ce cluster (i.e., satisfaits par la plupart de ses observations), discriminant par rapport aux autres clusters et concis. Notre méthode permet de construire simultanément un clustering et les explications de chaque cluster (*explanation by design*, comme par exemple dans (De Raedt et Blockeel, 1997; Dao et al., 2018; Fraiman et al., 2013)). Elle est inspirée par les méthodes de clustering ensemble (un ensemble de clusters est tout d'abord construit), et s'appuie sur la fouille de motifs (*pattern mining*) pour construire un ensemble de motifs pour chaque cluster. La dernière étape hautement combinatoire, implantée en Programmation Par Contraintes (PPC), sélectionne un ensemble de clusters et un ensemble de motifs pour chaque cluster, tout en respectant les contraintes utilisateur. Nos contributions sont donc les suivantes :

- un cadre générique pour le clustering explicable fondé sur deux espaces de représentation (clustering / explication),
- une nouvelle méthode de clustering explicable,
- un modèle en PPC pour la sélection des clusters et des explications,
- des expérimentations sur des jeux de données synthétiques et réelles dans le domaine de l'éducation (Cleuziou et Flouvat, 2021).

Le papier est organisé comme suit : état de l'art (Sec. 2), notre formulation du clustering explicable (Sec. 3), présentation de la méthode (Sec. 4), modèle en PPC (Sec. 5), résultats expérimentaux (Sec. 6) et conclusion (Sec. 7).

2 Travaux connexes

De nombreux travaux en clustering explicable s'appuient sur des arbres de décision. Certains de ces arbres sont construits a posteriori afin d'approcher les résultats obtenus par un algorithme de clustering (Dasgupta et al., 2020; Frost et al., 2020; Laber et al., 2023; Makarychev et Shan, 2021). Des approches intrinsèquement explicables (*explainable-by-design*) ont également été introduites avec des critères de séparation tels que la compacité des clusters (Loyola-Gonzalez et al., 2020) ou les indices de Silhouette et de Dunn (Bertsimas et al., 2021).

D'autres approches portent sur la délimitation des clusters à l'aide de formes englobantes : des hyper-rectangles (Chen et al., 2016) pour des contours parallèles aux axes, ou des polytopes (Lawless et al., 2022) pour des contours obliques.

Une troisième famille de techniques vise à intégrer dans un même cadre le clustering basé sur la distance et le clustering conceptuel (Dao et al., 2018; Davidson et al., 2018; Sambaturu et al., 2020; Zhang et Davidson, 2021). Elle a été introduite dans (Dao et al., 2018) sous le terme de clustering descriptif : les données sont décrites par deux modalités, l'une pour construire des clusters compacts et l'autre composée de descripteurs sémantiques pour former des descriptions interprétables. Nous introduisons un nouveau cadre de clustering descriptif qui à la différence de ces approches prend en compte des explications plus complexes sous la forme de motifs, c'est-à-dire de conjonctions de descripteurs booléens, et qui intègre la notion de discrimination entre clusters.

La plupart des approches de clustering sous contraintes construisent une partition par affection des points aux clusters, tout en satisfaisant les contraintes. Certains travaux génèrent d'abord un ensemble de partitions et sélectionnent ensuite la meilleure en fonction de sa capacité à satisfaire les contraintes (Van Craenendonck et Blockeel, 2017). À notre connaissance, seuls (Mueller et Kramer, 2010; Ouali et al., 2016; Chabert et Solnon, 2017) créent une partition en sélectionnant des clusters dans un ensemble donné de clusters tout en satisfaisant des contraintes définies par l'utilisateur. Par rapport à ces travaux, nous apportons la formalisation et l'intégration de la notion d'explicabilité pendant la construction d'un clustering, la prise en compte des contraintes de discrimination, le couplage du clustering à base de distance et du clustering conceptuel et enfin, inspirée par les méthodes ensemble, la prise en compte de diverses méthodes de clustering lors de la génération des partitions de base.

Notre approche, nommée ECS pour Explainability-driven Cluster Selection, s'inspire des méthodes ensemble qui reposent sur la génération d'un grand nombre de partitions de base et la construction d'une unique partition finale, souvent appelée partition consensus. Bien que diverses méthodes aient intégré des connaissances expertes dans le clustering ensemble (Guilbert et al., 2022; Yang et al., 2012, 2019; Yu et al., 2014, 2018), aucune ne s'est focalisée sur la sélection de clusters interprétables à partir d'un ensemble de clusters.

Nous nous différencions également du domaine appelé Redescription Mining (Galbrun et Miettinen, 2018), dont le but est de décrire un groupe d'objets dans un autre espace en maximisant l'accord entre les deux descriptions. Dans nos travaux, nous cherchons à structurer les données en clusters contenant des objets similaires, chacun étant associé à une explication, avec deux distinctions majeures : (1) les groupes ne sont pas connus à l'avance, (2) les explications d'un cluster dépendent des autres clusters.

Enfin, bien que les notions de couverture et de discrimination soient liées à l'apprentissage de concepts (voir par exemple (Michalski et Stepp, 1983) introduisant les notions de descriptions caractéristiques et discriminantes), notre travail s'en distingue dans la mesure où, dans notre approche, les classes et donc les concepts ne sont pas prédéfinis mais doivent être découverts.

3 Clustering explicable

Nous travaillons sur des données décrites par deux vues, possiblement construites à partir du même espace de représentation : une pour le clustering ($\mathbb{F}$) et une pour l'explication ($\mathbb{B}$). On

a donc un ensemble $\mathcal{O}$ de n objets, un ensemble de f caractéristiques (souvent numériques), un ensemble de r descripteurs booléens qui constituent les espaces $\mathbb{F}$ et $\mathbb{B}$, une matrice de taille $n \times f$ avec $\mathbb{F}_{oa}$ la valeur de l'objet o pour la propriété a et une matrice booléenne de taille $n \times r$ avec $\mathbb{B}_{op} = 1$, lorsque l'objet o satisfait la propriété p.

Un motif est une conjonction de descripteurs booléens et l'*explication* d'un cluster est un ensemble de motifs. Nous considérons qu'une explication d'un cluster C doit être *caractéristique* (couvrir la plupart des objets de C), *discriminante* (non satisfaite par la plupart des objets en dehors de C) et *concise* (les motifs ne doivent pas être trop longs et l'ensemble des motifs associés à un clustering ne doit pas être trop grand). Nous introduisons d'abord le prédicat $cover$ entre un motif π et un objet o par $cover(\pi, o) \stackrel{def}{=} \forall t \in \pi, \mathbb{B}_{o,t} = 1$.

Les propriétés attendues d'une explication sont alors les suivantes : (le symbole $\#$ indique la cardinalité d'un ensemble ; θ, θ', ϱ, ϕ sont des paramètres prédéfinis dans $[0, 1]$)

— Couverture d'un cluster C par un motif π :
$cover(\pi, C) \stackrel{def}{=} \#\{o \in C | cover(\pi, o)\} \geq \theta.\#C$

— Couverture d'un cluster C par un ensemble de motifs $\Pi = \{\pi_1, \ldots, \pi_j\}$:
$cover(\Pi, C) \stackrel{def}{=} \#\{o \in C | \exists l \in [1, j], cover(\pi_l, o)\} \geq \theta'.\#C$

— Discrimination d'un motif π :
  — par rapport au dataset : π ne doit pas couvrir plus d'un certain ratio ϱ des instances $o \notin C$: $disc\_global(\pi, C) \stackrel{def}{=} \#\{o \in \mathcal{O} \backslash C \mid cover(\pi, o)\} \leq \varrho \,.\#(\mathcal{O} \backslash C)$
  — par rapport à un cluster : pour tout cluster $C' \neq C$, π ne doit pas couvrir plus qu'un ratio ϕ des objets de C' :
  $disc\_cluster(\pi, C, C') \stackrel{def}{=} \#\{o \in C' \mid cover(\pi, o)\} \; \leq \phi \,.\#C'$

Les explications au sein d'un clustering sont souvent liées par une relation de généralité :

$$subsumes(\pi_1, \pi_2) \stackrel{def}{=} \pi_1 \subseteq \pi_2$$

Un motif plus général a une meilleure couverture à l'intérieur d'un cluster, mais un moins bon pouvoir discriminant. L'expert peut intégrer sa préférence par une contrainte stipulant que si π_1 et π_2 sont deux explications d'un même cluster et si $subsumes(\pi_1, \pi_2)$, π_1 ou π_2, selon sa préférence doit être conservée. D'autres contraintes exprimant des connaissances expertes peuvent également être intégrées ; pour une liste de ces contraintes, voir (Dao et Vrain, 2024).

4 Méthode proposée

Notre méthode vise à trouver un clustering satisfaisant des conditions à la fois sur la composition du clustering et sur les explications en termes de couverture et de discrimination, conformément à la section 3. La Figure 1 en présente les étapes. A la différence de méthodes existantes (Dao et al., 2015, 2018) qui construisent un clustering en recherchant à affecter des points aux clusters, notre approche s'inspire du clustering ensemble (Vega-Pons et Ruiz-Shulcloper, 2011) : elle construit un ensemble de clusters candidats et en sélectionne un sous-ensemble satisfaisant toutes les exigences. L'un des avantages est la maîtrise de la complexité : au lieu de trouver une affectation de n points aux clusters, le problème consiste à sélectionner k clusters parmi c clusters candidats, c étant contrôlé, souvent plus petit que n. Le clustering est un problème NP-difficile (Hansen et Delattre, 1978) tout comme le clustering sous contraintes (Davidson et Ravi, 2007). Puisque le problème de couverture d'ensembles peut se réduire au

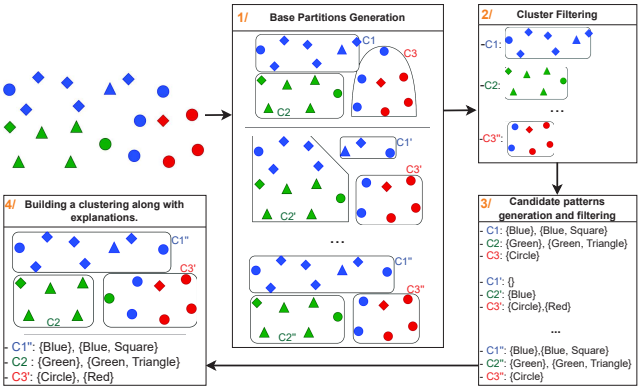

FIG. 1 – *Résumé de l'approche ECS.*

problème de sélection de clusters, ce dernier est NP-difficile. Les exigences sur le clustering et les explications sont formulées sous forme de contraintes, et réparties en trois catégories :

IC Contraintes sur la composition individuelle des clusters, telles que des restrictions sur leur cardinalité (nombre d'éléments), leur diamètre (distance maximale entre deux points dans un cluster), ou encore des contraintes sur les objets eux-mêmes, comme les contraintes must-link/cannot-link qui imposent que deux objets doivent ou ne doivent pas appartenir au même cluster.

ID Contraintes sur les explications des clusters indépendantes des autres clusters choisis, par exemple sur leur taille ou la couverture des motifs décrivant un cluster, ou sur la discrimination des motifs par rapport à l'ensemble des données.

AC Contraintes relatives à l'ensemble des clusters et motifs. Elles peuvent être basées sur l'explication (comme celles portant sur la capacité des motifs à discriminer d'autres clusters) ou sur le clustering, à l'instar de celle limitant le chevauchement entre clusters.

Construction des clusters candidats. En s'inspirant du clustering ensemble, divers algorithmes de clustering (K-means, clustering hiérarchique, clustering basé sur la densité) sont utilisés pour engendrer des partitions, dont les clusters constituent un ensemble de candidats (étape 1, Fig. 1). L'objectif est de garantir une diversité de formes parmi les clusters candidats : globulaires (K-means), allongés (clustering hiérarchique lien simple), denses (clustering fondé sur la densité). Les contraintes sur la composition individuelle des clusters **IC** sont ensuite appliquées, ne conservant que ceux qui les satisfont (étape 2).

Construction des motifs candidats. Pour chaque cluster candidat C, un ensemble de motifs candidats est généré. Ce sont les motifs fermés fréquents dans C générés via LCM (Uno et al., 2003) où la fréquence minimale est fixée par un seuil de couverture. La longueur des motifs peut être contrainte (un motif de taille 1 correspond alors à un seul descripteur). Les contraintes **ID** sont ensuite appliquées pour ne conserver que les motifs qui les respectent, ainsi que les clusters candidats possédant au moins un motif valide (étape 3).

Construction d'un clustering avec explications. Un clustering et l'explication de chaque cluster sont formés en sélectionnant des clusters parmi les candidats et, pour chaque cluster retenu, des motifs pour former son explication, tout en respectant les contraintes **AC** (étape 4).

Notons que la sélection des clusters et des motifs s'effectue simultanément. Ce problème est formulé comme un problème d'optimisation avec contraintes (COP) en PPC (section 5). Un COP se définit par un ensemble de variables de décision entières/booléennes et un ensemble de contraintes. Le solveur PPC cherche la meilleure solution en combinant propagation de contraintes et énumération (Rossi et al., 2006).

5 Modèle de sélection de clusters et de leurs explications

Soient $\mathcal{O}$ l'ensemble des N points et $\mathbf{C}$ l'ensemble des clusters candidats où chaque cluster $c \in \mathbf{C}$ est associé à un ensemble $\mathbf{D}_c$ de motifs candidats. L'objectif est de sélectionner un sous-ensemble de clusters dans $\mathbf{C}$ de sorte que les contraintes **AC** de clustering et de motifs soient satisfaites tout en optimisant une fonction objectif représentant la qualité du clustering et/ou de l'explication. Le problème a été formalisé en programmation par contraintes permettant d'exprimer, plus naturellement que la programmation linéaire en nombres entiers, des contraintes sous forme d'implications logiques et d'ajouter des contraintes non linéaires sur les clusters.

Le problème est modélisé avec les variables suivantes :

— Pour chaque cluster $c \in \mathbf{C}$, $C_c \in \{0,1\}$ une variable booléenne exprimant si le cluster c est sélectionné ;

— Pour chaque cluster $c \in \mathbf{C}$, pour chaque motif $d \in \mathbf{D}_c$, $D_{cd} \in \{0,1\}$ une variable exprimant si le motif d est sélectionné pour décrire c ;

— Pour chaque instance $x \in \mathcal{O}$, $A_x \in \mathbb{N}$ une variable entière comptant le nombre de clusters auquel x est affecté ;

— Pour chaque instance $x \in \mathcal{O}$, $P_x \in \{0,1\}$ une variable indiquant si x est expliqué par au moins un des motifs sélectionnés pour un cluster auquel il appartient.

Contraintes du modèle. Les variables A_x et P_x sont exprimées par les contraintes (avec $\mathbf{1}(\cdot)$ fonction indicatrice) :

$$\forall x \in \mathcal{O},\ A_x = \textstyle\sum_{c \in \mathbf{C}} C_c \mathbf{1}(x \in c)$$
$$\forall x \in \mathcal{O},\ P_x = \textstyle\sum_{c \in \mathbf{C}, d \in \mathbf{D}_c} D_{cd} \mathbf{1}(cover(d,x) \wedge x \in c)$$

Les contraintes de l'expert (Section 4) sont exprimées en PPC comme suit.

Contraintes sur la taille des explications.

- *Explication non-vide pour tout cluster sélectionné* : $\forall c \in \mathbf{C},\ C_c = 1 \iff \sum_{d \in \mathbf{D}_c} D_{cd} \geq 1$

En conséquence la non sélection d'un cluster entraîne la non sélection de ses motifs candidats.

- *Concision de l'explication (Miller, 1956)* (paramètre modifiable) : $\forall c \in \mathbf{C},\ \sum_{d \in \mathbf{D}_c} D_{cd} \leq 5$

Contraintes de discrimination.

- *Pour un motif, limitation du nombre de clusters acceptant ce motif comme explication.*

$\forall c \in \mathbf{C}, \forall d \in \mathbf{D}_c$,

$$D_{cd} = 1 \Longrightarrow \textstyle\sum_{c' \in \mathbf{C}} D_{c'd} \ \leq\ \eta \sum_{c' \in \mathbf{C}} C_{c'}$$

- *Un motif sélectionné pour un cluster c ne doit pas couvrir plus qu'un ratio ϕ des objets d'un autre cluster sélectionné c'.* $\forall c \in \mathbf{C}, \forall d \in \mathbf{D}_c$,

$$D_{cd} = 1 \Longrightarrow \textstyle\bigwedge_{c' \in \mathbf{C}, c' \neq c,\ freq(d,c') \geq \phi \# c'} (C_{c'} = 0)$$

Contraintes de couverture des explications.

Au moins un ratio α des points sont couverts par un motif : $\alpha n \leq \sum_{x \in \mathcal{O}} P_x$

Contraintes de clustering.
- Limitation du nombre de clusters sélectionnés : $K_{min} \leq \sum_{c \in \mathbf{C}} C_c \leq K_{max}$
- Limitation du nombre de clusters pour chaque instance :

$$\forall x \in \mathcal{O},\ \textit{nOverlap}_{min} \leq A_x \leq \textit{nOverlap}_{max}$$

- Limitation du chevauchement global des clusters : $\sum_{x \in \mathcal{O}} \mathbf{1}(A_x > 1) \leq$ *nOverlapAll*
Autres Contraintes. Des contraintes entre clusters sélectionnés peuvent être introduites, par exemple deux clusters c, c' ne peuvent pas être sélectionnés ensemble : $C_c + C_{c'} < 2$.

Objectifs du modèle. Plusieurs fonctions objectifs peuvent être utilisées, comme la minimisation de la somme des erreurs quadratiques ou le diamètre du cluster. Nous avons introduit de nouveaux objectifs : minimisation du nombre de points non affectés, maximisation du nombre de points affectés à un unique cluster, maximisation du nombre de points couverts par au moins un motif sélectionné. Dans les expérimentations nous considérons un objectif visant à affecter et expliquer un maximum de points tout en limitant les chevauchements défini par :

$$\arg\max \sum_{x \in \mathcal{O}} (P_x + \mathbf{1}(A_x = 1)) \tag{1}$$

6 Expérimentations

Dans cette section nous présentons des exemples concrets d'application à deux cas d'utilisation (éducation et automobile), ainsi qu'une comparaison avec des approches concurrentes.

6.1 Contexte expérimental

Nous nous intéressons à la qualité des explications obtenues et nous introduisons trois nouvelles mesures d'évaluation pour les explications dont les valeurs sont comprises entre 0 et 1 (1 le meilleur). Pour un motif p et un cluster C, PCR (resp. IPC) mesure le ratio d'observations couvertes (resp. non couvertes) par p dans (resp. en dehors de) C. Pour chaque cluster, on calcule d'abord la moyenne des PCR des motifs apparaissant dans son explication ; les tables donnent les moyennes de ces valeurs sur l'ensemble des clusters. Pour une explication D de C, EC mesure le nombre de points de C couverts par au moins un motif de D ; les tables donnent la moyenne d'EC calculée sur tous les clusters. Les mesures PCR et EC évaluent la couverture des explications et IPC la discrimination envers les autres clusters.

$$\text{Pattern Coverage Rate } \mathbf{PCR}(p, C) : \frac{\#\{o \in C : cover(p,o)\}}{\#C}$$

$$\text{Explanation Coverage } \mathbf{EC}(D, C) : \frac{\#\{o \in C | \exists p \in D\ cover(p,o)\}}{\#C}$$

$$\text{Inverse Pattern Contrastivity } \mathbf{IPC}(p, C) : \frac{1}{K-1} \sum_{C' \neq C} 1 - \frac{\#\{o \in C' | cover(p,o)\}}{\#C'}$$

L'approche est implantée en Python3 [1] ; les motifs sont générés avec l'implantation Python de LCM (Pedregosa et al., 2011) et le modèle PPC a été développé avec CPMPY (Guns, 2019). Les expériences ont été réalisées sur un ordinateur équipé d'un processeur 11th Gen Intel® Core™ i7-1165G7 @ 2,80GHz × 8 et d'une mémoire de 31,0 Go. Dans les tableaux, les motifs sont écrits entre accolades et séparés par ||. Le caractère - indique qu'aucune solution n'a été trouvée. Des couleurs différentes sont utilisées pour faciliter la lisibilité.

1. https ://github.com/MathieuGuilbert/ECS

C	**Explanation**	Size	PCR	EC	IPC
0	{user4} || {user12}	24	0.35	0.71	0.99
1	{user16}|| {user13}	14	0.46	0.93	1.0
2	{user39}	15	0.73	0.73	1.0
3	{user19}	6	1.0	1.0	1.0
4	{user28}	15	0.33	0.33	0.99
5	{user50}	7	0.71	0.71	1.0
6	{user36}	19	0.37	0.37	1.0
7	{user18}	5	1.0	1.0	1.0

TAB. 1 – *NC-1014, premier exercice, des motifs de longueur 1,* $\theta = 0.5, \phi = 0.3$.

6.2 Résultats expérimentaux

6.2.1 Cas d'étude dans l'éducation

NC-1014 est composé de 1014 programmes écrits en 2020 par 60 étudiants de l'Université de Nouvelle-Calédonie répondant à 8 exercices différents d'un module d'apprentissage Python. L'espace de propriétés pour le clustering $\mathbb{F}$ est un embedding de dimension 20 calculé à l'aide de la méthode *sec2vec* (Martinet et al., 2024) sur les programmes bruts. Aucune information n'a été donnée au modèle à propos des auteurs des programmes, ni sur les exercices qu'ils sont censés résoudre. L'espace d'explication $\mathbb{B}$ est composé de descripteurs indiquant la présence de tokens importants tels que *if, else, while, for*, le numéro de l'exercice, l'utilisateur, et certaines valeurs numériques discrétisées en fonction de la médiane (par exemple le nombre de retours).

NC-1014 - jeu de donnée complet. L'objectif est de maximiser le nombre de points couverts et attribués à un seul cluster (Eq. 1). Nous avons observé que la plupart des clusters sont partiellement expliqués par un exercice, ce qui suggère que les programmes répondant au même exercice ont des embeddings similaires. Chaque cluster a en effet dans son explication au moins un motif avec des informations sur la présence d'un exercice et les propriétés communes des codes donnés en réponse. Par exemple, un cluster est constitué majoritairement de réponses à l'exercice 6 contenant des boucles imbriquées et des fonctions avec plusieurs paramètres.

NC-1014 - premier exercice. La méthode a été appliquée aux programmes correspondant au premier exercice avec des motifs de longueur 1 (un seul descripteur) et avec des motifs de longueur quelconque. Le tableau 1 montre que les clusters correspondent globalement à des étudiants différents. Les résultats diffèrent quelque peu lorsque des motifs de longueur variable sont utilisés, les explications intégrant alors des informations sur les étudiants et leur style de codage.

Par exemple, le cluster principalement composé de réponses soumises par l'étudiant 39 est caractérisé par des programmes contenant plus de 10 lignes, des fonctions avec plusieurs paramètres, des boucles while, des commentaires, ...Rappelons que l'identité des auteurs des programmes n'a pas été incluse lors du calcul de l'embedding. Cela suggère que le style d'écriture y est induit. Ainsi, même lorsque des clusters sont générés dans l'espace d'embedding, ils peuvent être expliqués par le style de codage de certains étudiants.

6.2.2 Jeu de données Automobile

Le jeu de données UCI Automobile (Dua et Graff, 2017) décrit 150 voitures (après suppression de 55 observations incomplètes) par 24 propriétés descriptives et leur prix. Le par-

C	**Explanation**	Size	EC	IPC	Price
0	{mpfi-fuel-injection}	42	0.88	0.68	17090
1	{turbo} || {five-cylinders} || {mercedes-benz}	7	0.86	0.93	29251
2	{curb-weight $\leq$ 2340} || {city-mpg $>$ 26} || {highway-mpg $>$ 32} || {front-wheel-traction}	110	0.91	0.93	8157

TAB. 2 – *Automobile, longueur de motif 1, K=3,* $\theta = 0.7$*,* $\phi = 0.5$.

titionnement a été effectué sur le prix et les clusters expliqués par les autres propriétés. Les attributs numériques sont discrétisés selon la médiane. Les partitions de base (étape 1) ont été générées par K-means et le clustering spectral avec un nombre de clusters variant de 2 à 15. Le tableau 2 présente les résultats avec des motifs de longueur 1. Le temps d'exécution du modèle en PPC dépend du nombre de motifs candidats. Plus le paramètre de couverture est élevé, plus le nombre de motifs candidats augmente et plus le temps d'exécution est important. Ainsi, les temps d'éxécution du modèle avec en paramètre $\theta = 0.9$ et $\phi = 0.5$ contre $\theta = 0.5$ et $\phi = 0.3$ passent d'environ 30 a 900 secondes. Pour cette raison, notre méthode est plus rapide lorsque qu'elle autorise uniquement l'utilisation de motifs de longueur 1. Nous remarquons que les clusters représentent 3 catégories de prix avec des voitures bon marché, moyennes et chères, chacune caractérisée par des propriétés intrinsèques. Par exemple, le cluster 1 est composé des 7 voitures les plus chères caractérisées par un moteur turbo à cinq cylindres et/ou par la marque Mercedes-Benz. Le cluster 0 est lui caractérisé par la présence de systèmes d'injection MFPI.

6.2.3 Comparaison avec d'autres travaux

Nous considérons deux autres approches : la première applique d'abord l'algorithme K-Means puis recherche pour chaque cluster des motifs de longueur 1 (*K-Means*) ou des motifs de longueur quelconque (*K-Means–LCM*) ; la seconde présentée dans (Dao et al., 2018) ne considère que des motifs de longueur 1 (descripteur unique). Cette dernière vise à trouver un front de Pareto de partitions maximisant le diamètre des clusters et un critère MMCTA comptant le nombre de descripteurs partagés par toutes les instances d'un cluster. Nous comparons nos résultats avec 3 résultats issus de leur front de Pareto : le premier maximisant MMCTA, le dernier maximisant le diamètre des clusters et le médian (noté *central solution*).

Nos expériences, présentées dans les tables 3 et 4, ont été menées sur les ensembles de données *Automobile*, *NC-1014* et UCI-*Student Performance* (Cortez et Silva, 2008).

Ces résultats illustrent la capacité d'ECS à générer des explications plus discriminantes que celles de nos compétiteurs permettant ainsi aux utilisateurs de mieux saisir les différences entre les objets des différents clusters.

7 Conclusion

Nous proposons une nouvelle méthode de clustering *explicable par design*, où une explication consiste en un ensemble de motifs couvrant la majorité des instances d'un cluster et les discriminant des autres. Ces notions sont exprimées par des contraintes, permettant à l'expert d'intégrer ses connaissances et attentes. Deux espaces de représentation sont utilisés : l'un pour le clustering et un autre, booléen, pour les explications. Bien que l'espace d'explication puisse dériver de celui du clustering, nous avons démontré l'intérêt de cette séparation sur deux cas

	Dataset	Auto	Student	NC
	K	3	3	8
K-Means	PCR	0.87	0.83	0.89
	EC	1.0	1.0	1.0
	IPC	0.48	0.21	0.4
K-Means (LCM)	PCR	0.8	0.78	0.84
	EC	1.0	1.0	1.0
	IPC	0.77	0.28	0.63
ECS	PCR	0.86	-	0.83
	EC	0.94	-	0.83
	IPC	0.82	-	0,97
ECS (LCM)	PCR	0.81	0.74	0.83
	EC	0.99	0.89	0.83
	IPC	0.89	0.61	0.98

TAB. 3 – *Résultats : baseline K-means et ECS avec* $\theta = 0.7$ *et* $\phi = 0.5$.

	Dataset	Auto	Student	NC
	K	3	3	8
Dao *et al.* 2018, max MMCTA	PCR	1.0	1.0	1.0
	EC	1.0	1.0	1.0
	IPC	0.42	0.14	0,38
Dao *et al.* 2018, central solution	PCR	1.0		1.0
	EC	1.0		1.0
	IPC	0.25		0.25
Dao *et al.* 2018, max diameter	PCR	1.0	0.0	1.0
	EC	1.0	0.0	1.0
	IPC	0.19	0.0	0.15

TAB. 4 – *Résultats : Dao et al. 2018. NB : seulement 2 solutions dans le front de Pareto pour Student.*

d'usage (éducation et automobile). Nous envisageons l'ajout de nouvelles contraintes et critères objectifs et une application de la méthode à la chemo-informatique, où les molécules doivent être expliquées par des caractéristiques structurelles comme les propriétés pharmacophoriques, justifiant leurs activités biologiques.

Remerciements. Ce travail est financé par le projet ANR InvolvD (Interactive constraint elicitation for unsupervised and semi-supervised data mining) (ANR-20-CE23-0023).

Références

Bertsimas, D., A. Orfanoudaki, et H. Wiberg (2021). Interpretable clustering : an optimization approach. *Machine Learning 110*(1), 89–138.

Chabert, M. et C. Solnon (2017). Constraint programming for multi-criteria conceptual clustering. In *CP*, pp. 460–476.

Chen, J., Y. Chang, B. Hobbs, P. Castaldi, M. Cho, E. Silverman, et J. Dy (2016). Interpretable clustering via discriminative rectangle mixture model. In *ICDM*, pp. 823–828. IEEE.

Cleuziou, G. et F. Flouvat (2021). Learning student program embeddings using abstract execution traces. In *14th Int. Conf. on Educ. DM*, pp. 252–262.

Cortez, P. et A. M. G. Silva (2008). Using data mining to predict secondary school student performance.

Dao, T. et C. Vrain (2024). A review on declarative approaches for constrained clustering. *Int. J. Approx. Reason. 171*, 109–135.

Dao, T.-B.-H., C.-T. Kuo, S. Ravi, C. Vrain, et I. Davidson (2018). Descriptive clustering : Ilp and cp formulations with applications. In *27th IJCAI*, pp. 1263–1269.

Dao, T.-B.-H., W. Lesaint, et C. Vrain (2015). Clustering conceptuel et relationnel en programmation par contraintes. *JFPC 11*.

Dasgupta, S., N. Frost, M. Moshkovitz, et C. Rashtchian (2020). Explainable k-means and k-medians clustering. In *ICML*, pp. 12–18.

Davidson, I., A. Gourru, et S. Ravi (2018). The cluster description problem-complexity results, formulations and approximations. *Advances in Neural Information Processing Systems 31*.

Davidson, I. et S. S. Ravi (2007). The Complexity of Non-hierarchical Clustering with Instance and Cluster Level Constraints. *Data Mining Knowledge Discovery*.

De Raedt, L. et H. Blockeel (1997). Using logical decision trees for clustering. In *International Conference on Inductive Logic Programming*, pp. 133–140. Springer.

Dua, D. et C. Graff (2017). UCI machine learning repository.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Mach. Learn. 2*(2), 139–172.

Fraiman, R., B. Ghattas, et M. Svarc (2013). Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification 7*(2), 125–145.

Frost, N., M. Moshkovitz, et C. Rashtchian (2020). Exkmc : Expanding explainable k-means clustering. *arXiv preprint arXiv :2006.02399*.

Galbrun, E. et P. Miettinen (2018). *Redescription Mining* (1st ed.). Springer Publishing Company, Incorporated.

Guilbert, M., C. Vrain, M. C. de Souto, et al. (2022). Anchored constrained clustering ensemble. In *IJCNN*, pp. 1–8. IEEE.

Guns, T. (2019). Increasing modeling language convenience with a universal n-dimensional array, cppy as python-embedded example. In *18th Modref workshop at CP*.

Hansen, P. et M. Delattre (1978). Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association 73*(362), 397–403.

Laber, E., L. Murtinho, et F. Oliveira (2023). Shallow decision trees for explainable k-means clustering. *Pattern Recognition 137*, 109239.

Lawless, C., J. Kalagnanam, L. M. Nguyen, D. Phan, et C. Reddy (2022). Interpretable clustering via multi-polytope machines. In *AAAI*, Volume 36, pp. 7309–7316.

Loyola-Gonzalez, O., A. E. Gutierrez-Rodríguez, M. A. Medina-Pérez, R. Monroy, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, et M. Garcia-Borroto (2020). An explainable artificial intelligence model for clustering numerical databases. *IEEE Access 8*, 52370–52384.

Makarychev, K. et L. Shan (2021). Near-optimal algorithms for explainable k-medians and k-means. In *ICML*, pp. 7358–7367.

Martinet, T., G. Cleuziou, M. Exbrayat, et F. Flouvat (2024). From document to program embeddings : can distributional hypothesis really be used on programming languages ? In *ECAI*.

Michalski, R. S. et R. E. Stepp (1983). Automated construction of classifications : Conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*.

Miller, G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *Psychological review 63*(2), 81.

Mueller, M. et S. Kramer (2010). Integer linear programming models for constrained clustering. In *DS*, pp. 159–173.

Ouali, A., S. Loudni, Y. Lebbah, P. Boizumault, A. Zimmermann, et L. Loukil (2016). Efficiently finding conceptual clustering models with integer linear programming. In *IJCAI*, pp. 647–654.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn : Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Rossi, F., P. van Beek, et T. Walsh (2006). *Handbook of Constraint Programming*. Foundations of Artificial Intelligence. Elsevier B.V.

Sambaturu, P., A. Gupta, I. Davidson, S. Ravi, A. Vullikanti, et A. Warren (2020). Efficient algorithms for generating provably near-optimal cluster descriptors for explainability. In *AAAI*, Volume 34, pp. 1636–1643.

Uno, T., T. Asai, Y. Uchida, et H. Arimura (2003). Lcm : An efficient algorithm for enumerating frequent closed item sets. In *Fimi*, Volume 90.

Van Craenendonck, T. et H. Blockeel (2017). Constraint-based clustering selection. *Machine Learning 106*, 1497–1521.

Vega-Pons, S. et J. Ruiz-Shulcloper (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence 25*(03), 337–372.

Yang, J., L. Sun, et Q. Wu (2019). Constraint projections for semi-supervised spectral clustering ensemble. *Concurrency and Computation : Practice and Experience*.

Yang, Y., H. Wang, C. Lin, et J. Zhang (2012). Semi-supervised clustering ensemble based on multi-ant colonies algorithm. In *RSKT*, pp. 302–309. Springer.

Yu, Z., H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, et G. Han (2014). Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 11*(4), 727–740.

Yu, Z., P. Luo, J. Liu, H.-S. Wong, J. You, G. Han, et J. Zhang (2018). Semi-supervised ensemble clustering based on selected constraint projection. *IEEE TKDE*.

Zhang, H. et I. Davidson (2021). Deep descriptive clustering. In *IJCAI*.

Summary

Clustering is an unsupervised exploratory task that helps experts understanding the structure of their data. Constraints based on their knowledge can be introduced, but obtaining them remains challenging, making the explanation of results essential for adjusting parameters and uncovering new insights. We address explainable clustering by modeling the data in two spaces: one for clustering and another for explanation. Our method ECS (Explainability-driven Cluster Selection) aims to produce a high-quality clustering while ensuring interpretability through patterns that cover most instances in a cluster and distinguish them from others. It relies on ensemble clustering and a new constraint programming model for selecting the clusters and their explanations.