Vers la découverte de modèles exceptionnels locaux : des règles descriptives liant les molécules à leurs odeurs

Guillaume Bosc*, Mehdi Kaytoue*, Marc Plantevit***,
Fabien De Marchi***, Moustafa Bensafi**, Jean-François Boulicaut*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, France
**Centre National de la Recherche Scientifique UMR5292, INSERM U1028,
CRNL, Université Claude Bernard Lyon 1 Lyon, France
***Université de Lyon, CNRS, Université Lyon 1 LIRIS UMR5205, F-69622, France
Contact : guillaume.bosc@insa-lyon.fr

Résumé. Issue d'un phénomène complexe partant d'une molécule odorante jusqu'à la perception dans le cerveau, l'olfaction reste le sens le plus difficile à appréhender par les neuroscientifiques. L'enjeu principal est d'établir des règles sur les propriétés physicochimiques des molécules (poids, nombre d'atomes, etc.) afin de caractériser spécifiquement un sous-ensemble de qualités olfactives (fruité, boisé, etc.). On peut trouver de telles règles descriptives grâce à la découverte de sous-groupes ("subgroup discovery"). Cependant les méthodes existantes permettent de caractériser soit une seule qualité olfactive ; soit toutes les qualités olfactives à la fois ("exceptional model mining") mais pas un sous-ensemble. Nous proposons alors une approche de découverte de sous-groupes caractéristiques de seulement certains labels, par une nouvelle technique d'énumération, issue de la fouille de redescriptions. Nous avons expérimenté notre méthode sur une base de données d'olfaction fournie par des neuroscientifiques et pu exhiber des premiers sous-groupes intelligibles et réalistes.

1 Introduction

L'olfaction, ou la capacité de percevoir des odeurs, est le résultat d'un phénomène complexe : une molécule s'associe à un récepteur de la cavité nasale, et provoque l'émission d'un signal transmis au cerveau qui fait ressentir l'odeur associée [Sezille et Bensafi (2013)–Meierhenrich et al. (2005)]. Si les phénomènes qui caractérisent les sens de l'ouïe et de la vue sont bien connus, la perception olfactive n'est, encore aujourd'hui, toujours pas comprise dans sa globalité. Cependant, on dispose de nombreux atlas (comme celui d'Arctander (1969)) qui renseignent les qualités perçues par l'humain pour des milliers de molécules odorantes : des experts senteurs associent à des milliers de molécules odorantes des qualités d'odeurs (fruité, boisé, huileux, etc : un vocabulaire bien défini et consensuel). On dispose également maintenant d'outils capables de calculer des milliers de propriétés physico-chimiques de molécules [1]. Il a alors pu être montré que ces propriétés déterminent la (les) qualité(s) d'une odeur perçue

1. Par exemple Dragon 6 – `http://www.talete.mi.it/`

[Khan et al. (2007)–Kaeppler et Mueller (2013)]. Ce lien entre le monde physico-chimique et le monde du percept olfactif a été mis en évidence à l'aide de méthodes d'analyse en composantes principales démontrant, à partir de données, la corrélation existante entre ces deux mondes. Les neuroscientifiques ont donc maintenant besoin de méthodes descriptives afin de comprendre les liens entre propriétés physicochimiques et qualités.

La découverte de régularités (ou descriptions) qui distinguent un groupe d'objets selon un label cible (souvent appelé label de classe), est un problème qui a fédéré diverses communautés en intelligence artificielle, fouille de données, apprentissage statistique, etc. En particulier, la découverte supervisée de règles descriptives de type $description \longrightarrow label$ est étudiée sous divers formalismes : découverte de sous-groupes, fouille de motifs émergents, ensembles contrastés, hypothèses, etc. (Novak et al. (2009)). Dans tous les cas, nous faisons face à un ensemble d'objets associés à des descriptions (dont l'ensemble forme un ensemble partiellement ordonné), et ces objets sont liés à un ou plusieurs labels de classe.

Dans cet article, on s'intéresse à la découverte de sous-groupes (subgroup discovery), introduite par Klösgen (1996) et Wrobel (1997). Étant donné un ensemble d'objets décrits par un ensemble d'attributs, et chacun associé à un (ou plusieurs) label(s) de classe, un sous-groupe est un sous-ensemble d'objets statistiquement intéressant par sa taille et ses singularités au sein de l'ensemble d'objets initial vis à vis d'un ou plusieurs labels cibles. En fait, il existe deux familles principales de méthodes. La première (Wrobel, 1997) vise à trouver des règles de type $description \rightarrow label$ où le conséquent est un unique label. La seconde, la fouille de modèles exceptionnels (exceptional model mining, EMM) introduite par Leman et al. (2008), vise à trouver des sous-groupes dont la répartition d'apparition de *tous* les labels diffèrent grandement dans le sous-groupe comparé à toute la population, i.e. de la forme $description \rightarrow \{(label_1, valeur_1), ..., (label_k, valeur_k)\}$ où k est le nombre de labels de l'attribut cible. Dans les deux cas, on veut optimiser une mesure de qualité pour distinguer au mieux le sous-groupe en fonction du label, ou d'une distribution des labels dans le sous-groupe (i.e. le modèle).

En olfaction cependant, une molécule est associée à une ou plusieurs qualités d'odeurs : aucune des approches existantes ne permet de se focaliser sur un sous-ensemble de labels de cardinalité arbitraire. Effectivement, ces approches permettent soit de caractériser un seul label de classe par des sous-groupes, soit de trouver des sous-groupes qui caractérisent tous les labels de classes à la fois. Alors, d'une part, un sous-groupe effectue une caractérisation trop locale, trop spécifique et d'autre part la caractérisation est beaucoup trop globale.

Nous cherchons alors à découvrir des sous-groupes comme des règles descriptives de type $description \rightarrow \{label_1, label_2, ..., label_l\}$ où $l << k$. Pour cela, nous proposons une nouvelle méthode appelée ElMM (Exceptional local Model Mining) qui généralise à la fois la méthode de sous-groupes classiques ainsi que EMM. Nous montrerons alors que les sous-groupes extraits sont plus caractéristiques de peu de qualités à la fois, et donc aussi plus faciles à interpréter par l'expert en olfaction.

La suite de cet article est organisée comme suit. Tout d'abord, nous introduisons les deux principales méthodes de découverte de sous-groupes en section 2 (*subgroup discovery* et *exceptionnal model mining*). Nous montrons alors les limites de ces deux types d'approche avant d'introduire en section 3 notre nouvelle méthode : la découverte de modèles exceptionnels locaux (*exceptional local model mining)*. Un algorithme de découverte est présenté en section 4 et appliqué à des données issues des domaines de la neuroscience et de l'olfaction (section 5).

2 Découverte de sous-groupes : définition et problème

Nous présentons tout d'abord les notations utilisées dans cet article, les méthodes de découverte de sous-groupes et de fouille de modèles exceptionnels ainsi que leur limite.

Définition 1 (Jeu de données). Soient $\mathcal{O}$ un ensemble d'objets et $\mathcal{A}$ un ensemble d'attributs. Un attribut $a \in \mathcal{A}$ est nominal quand son domaine est de la forme $Dom(a) = \{a_1, a_2, ..., a_m\}$, et numérique quand $Dom(a)$ est un sous ensemble fini de nombres réels (correspondant aux valeurs prises par les objets pour a). On distingue de plus un attribut nominal de classe C, et la fonction $class : \mathcal{O} \mapsto 2^{Dom(C)}$ qui associe à tout objet de $\mathcal{O}$ une ou plusieurs valeurs (labels) du domaine de l'attribut de classe. On note $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ un tel jeu de données.

Exemple. Soit un ensemble de molécules $\mathcal{O}$, identifiées par leur identifiant $\{1, 24, 48, 60, 82, 1633\}$, décrites par un ensemble de propriétés physicochimiques $\mathcal{A}$ (le poids moléculaire (MW), le nombre d'atomes (nAT) et le nombre d'atomes de carbone (nC)) et associées à leur qualité olfactive C parmi $\{Fruité, Miellé, Vanillé\}$. La Table 1 (resp. 2) présente ce jeu de données dans le cas où la fonction $class$ n'associe à chaque molécule qu'un seul label de C –mono-qualité– (resp. un sous-ensemble de labels –multi-qualités–).

Un sous-groupe peut être représenté formellement de manière duale soit en intension soit en extension, c'est-à-dire, soit par une description dans un langage donné mettant en œuvre des restrictions sur le domaine de valeurs des attributs, soit par l'ensemble d'objets qu'il décrit. Il existe plusieurs langages de description possibles, basés sur différents types de connecteurs logiques (conjonctions, disjonctions, ou encore négations), dont certains sont très expressifs (voir par exemple Galbrun et Kimmig (2014)). Dans la suite nous utiliserons un langage basé uniquement sur des conjonctions.

Définition 2 (Sous-groupe). On note $d = \langle f_1, \ldots, f_{|\mathcal{A}|} \rangle$ la description d'un sous-groupe où chaque f_i est une restriction sur le domaine de l'attribut $a_i \in \mathcal{A}$ (à un sous-ensemble du domaine de a_i s'il est nominal, ou à un intervalle s'il est numérique). Chaque restriction peut être assimilée soit à un ensemble (dans le cas d'une restriction sur un attribut nominal), soit à un intervalle dont les bornes appartiennent à $Dom(a_i)$ (dans le cas d'un attribut numérique). L'ensemble d'objets qui vérifient une description d est appelé support, noté $supp(d) \subseteq \mathcal{O}$.

Relation d'ordre partiel entre les sous-groupes. Soient $d_1 = \langle f_1^1, \ldots, f_{|\mathcal{A}|}^1 \rangle$ et $d_2 = \langle f_1^2, \ldots, f_{|\mathcal{A}|}^2 \rangle$ deux descriptions relatives à deux sous-groupes différents, où chaque restriction f_i^j, avec $i \in \{1, 2, ..., |\mathcal{A}|\}$ et $j \in \{1, 2\}$, est la restriction sur l'attribut $a_i \in \mathcal{A}$, alors

ID	MW	nAT	nC	Qualité
1	150.19	21	11	{Fruité}
24	128.24	29	9	{Vanillé}
48	136.16	24	10	{Miellé}
60	152.16	23	11	{Fruité}
82	151.28	27	12	{Miellé}
1633	142.22	27	10	{Fruité}

TAB. 1: Base de molécules mono-qualité

ID	MW	nAT	nC	Qualité
1	150.19	21	11	{Fruité}
24	128.24	29	9	{Miellé, Vanillé}
48	136.16	24	10	{Miellé, Vanillé}
60	152.16	23	11	{Fruité}
82	151.28	27	12	{Miellé, Fruité}
1633	142.22	27	10	{Fruité, Vanillé}

TAB. 2: Base de molécules multi-qualités

$f_i^j = [x, y]$, avec $x, y \in Dom(a_i)$, si a_i est numérique, et $f_i^j \subseteq Dom(a_i)$, si a_i est nominal. On dit alors que d_1 est une spécialisation de d_2, et on le note $d_1 \preceq d_2$, si et seulement si pour tout $i \in \{1, 2, ..., |\mathcal{A}|\}$, on a $f_i^1 \subseteq f_i^2$ si a_i est numérique et $f_i^1 \supseteq f_i^2$ si a_i est nominal.

Exemple (suite). On a $d_1 = \langle MW \leq 151.28, 23 \leq nAT \rangle$ avec pour support l'ensemble des molécules $\{24, 48, 82, 1633\}$: la molécule 1 ne vérifie pas la restriction sur l'attribut nAT alors que la molécule 60 ne vérifie pas celle sur MW. Pour plus de lisibilité, lorsque l'on ne précise pas une restriction f_i dans une description d cela signifie qu'aucune restriction effective n'est appliquée sur l'attribut a_i dans d. La description $d_2 = \langle MW \leq 151.28, 23 \leq nAT, 10 \leq nC \rangle$ est une spécialisation de d_1 car d_2 comporte les mêmes restrictions que d_1 plus une restriction sur un autre attribut. Réciproquement, d_1 est une généralisation de d_2.

Étant donné un jeu de données, il y a potentiellement $2^{|\mathcal{O}|}$ sous-groupes, il est donc nécessaire de n'en sélectionner qu'une partie en fonction de leur intérêt. Pour cela, les différentes approches de l'état de l'art utilisent une mesure de qualité qui évalue la singularité du groupe au sein de la population par rapport à une cible, c'est-à-dire l'attribut de classe. La mesure de qualité est choisie en fonction du type de données, mais aussi en fonction de l'attribut de classe et de la finalité de l'application. Il existe deux approches pour la découverte de sous-groupes : l'approche que l'on va définir comme classique (Wrobel, 1997), et l'approche d'Exceptional Model Mining (EMM) introduite par Leman et al. (2008).

Dans la première, chaque objet n'est associé qu'à un et un seul label de l'attribut de classe, c'est-à-dire $\forall o \in \mathcal{O}$, $class(o) = c$ avec $c \in Dom(C)$, et la mesure de qualité permet de mettre en évidence la singularité d'un sous-groupe relativement à un seul label de C. Pour un sous-groupe de description d, une mesure généralement utilisée relativement au label l est :

$$M(d, l) = \left(\frac{|supp(d)|}{|\mathcal{O}|} \right)^{\alpha} \times (p^l - p_0^l) \tag{1}$$

avec $0 \leq \alpha \leq 1$, $p^l = \frac{|\{o \in supp(d) | class(o) = l\}|}{|supp(d)|}$ et $p_0^l = \frac{|\{o \in \mathcal{O} | class(o) = l\}|}{|\mathcal{O}|}$ les proportions d'objets du sous-groupe et du jeu de données entier possédant la classe l. Cette mesure est une généralisation de la mesure *WRAcc* ($\alpha = 1$) qui prend en compte à la fois la taille du sous-groupe et aussi sa singularité dont le rapport entre les deux est pondéré par un facteur α.

Dans le cas d'EMM, un objet est associé à un sous-ensemble de labels de classe, c'est-à-dire, $\forall o \in \mathcal{O}$, $class(o) \subseteq Dom(C)$. La mesure de qualité utilisée dans ce cas permet de mettre en évidence la singularité d'un sous-groupe relativement à tous les labels de C à la fois. Une mesure possible est la somme des divergences de Kullback-Leibler pour tous les labels de classe entre les objets du sous-groupe et ceux du jeu de données entier :

$$WKL(d) = \frac{|supp(d)|}{|\mathcal{O}|} \sum_{l \in Dom(C)} (p^l \log_2 \frac{p^l}{p_0^l}) \tag{2}$$

avec $p^l = \frac{|\{o \in supp(d) | l \in class(o)\}|}{|supp(d)|}$ et $p_0^l = \frac{|\{o \in \mathcal{O} | l \in class(o)\}|}{|\mathcal{O}|}$

Exemple (suite). Avec la description $d_1 = \langle MW \leq 151.28, 23 \leq nAT \rangle$, dans la Table 1, en utilisant la mesure de l'équation (1) avec $l = Miellé$ et $\alpha = 1$, on obtient $M(d_1, l) = 4/6 \times (1/2 - 1/3) = 2/3$. Dans la Table 2, en utilisant la mesure de la formule de l'équation (2), on a $WKL(d_1) = 4/6 \times ((2/4 \log_2 3/4) + (3/4 \log_2 3/2) + (3/4 \log_2 3/2)) = 0.45$.

Découverte de sous-groupes, limites et problème. Étant donnés $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$, $minSupp$, φ et k l'objectif est de récupérer l'ensemble des k-meilleurs sous-groupes au regard de la mesure de qualité φ choisie où la taille du support du sous-groupe est supérieure ou égale à $minSupp$. Pour notre domaine d'application de l'olfaction, les approches existantes (découverte de sous-groupes classique et EMM) ne permettent pas de répondre à la problématique posée, à savoir la caractérisation de sous-ensemble de qualités d'odeurs. Effectivement, ces approches permettent soit de caractériser un seul label de classe, c'est-à-dire une qualité olfactive, par sous-groupe, soit de trouver des sous-groupes qui caractérisent tous les labels de classes à la fois avec EMM. D'une part un sous-groupe effectue une caractérisation trop locale et spécifique, d'autre part la caractérisation est trop globale. Nous introduisons dans la suite une nouvelle méthode qui généralise ces deux approches en permettant de caractériser par un sous-groupe un sous-ensemble L de taille quelconque de labels de classe.

3 Découverte de modèles exceptionnels locaux : ElMM

Soit $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ un jeu de données conforme à la Définition 1, avec $Dom(C) = \{l_1, \ldots, l_k\}$. Étant donnée une mesure de qualité φ, notre méthode ElMM recherche des sous-groupes de la forme (d, L) où d est la description d'un sous-groupe et $L \subseteq Dom(C)$ est un sous-ensemble de labels de la classe C à caractériser. Cette méthode correspond au cas général de la découverte de sous-groupes. Effectivement, si on fixe pour tous les sous-groupes que $L \in Dom(C)$ on se ramène au cas de la découverte de sous-groupes classique dans lequel un sous-groupe ne caractérise qu'un label de classe. De plus si $L = Dom(C)$ alors on bascule dans le cas d'EMM où chaque sous-groupe doit caractériser tous les labels de classe à la fois. ElMM permet donc de caractériser des sous-ensembles de labels de classe par des sous-groupes appelés sous-groupes locaux.

Définition 3 (Sous-groupe local). Soit $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ un jeu de données, on appelle sous-groupe local, et on le note (d, L), un sous-groupe de description d caractérisant un sous-ensemble de labels $L \subseteq Dom(C)$ de la classe C. De plus, un sous-groupe $S = (d, L)$ doit vérifier les trois contraintes suivantes : soient $minSupp, maxDesc$, et $maxLab \in \mathbb{N}^+$, (i) $|supp(d)| \geq minSupp$, (ii) $|d| \leq maxDesc$ et (iii) $|L| \leq maxLab$.

La contrainte (i) permet de ne considérer que les sous-groupes dont le support est supérieur ou égal à un seuil $minSupp$, évitant ainsi d'obtenir des sous-groupes de trop petite taille qui n'auraient alors aucun intérêt et facilitant l'exploration. La contrainte (ii) permet d'interagir sur le langage de la description en restreignant le nombre maximal de restrictions effectives par description à un seuil $maxDesc$ ($|d|$ est le nombre de restrictions effectives de d). De manière similaire, la contrainte (iii) permet de limiter le nombre de labels à discriminer dans L.

Mesure de qualité. La mesure de qualité utilisée dans EMM dont la formule a été donnée dans l'équation 2 peut être généralisée pour ElMM pour ne considérer qu'un sous-ensemble $L \subseteq Dom(C)$ de labels de classe et non plus l'ensemble complet de labels à la fois :

$$WKLGen(d, L) = \frac{|supp(d)|}{|\mathcal{O}|} \sum_{l \in L} (p^l \log_2 \frac{p^l}{p_0^l}).$$

Cependant, cette mesure de qualité ne correspond pas à l'objectif de notre contexte applicatif car elle ne quantifie pas les labels de L ensemble, c'est-à-dire de manière conjointe, lorsqu'ils sont associés conjointement aux objets. Nous cherchons à caractériser l'ensemble des objets cohérents qui possèdent tous les labels de L, et non pas un sous-ensemble de L. Pour cela, nous nous sommes tournés vers une mesure de qualité usuellement utilisée en classification supervisée : la F_1-*Mesure*. Cette mesure nous permet dans notre cas de quantifier la pureté d'un sous-groupe vis à vis des labels à caractériser L, i.e. les objets du support de la description du sous-groupe doivent être le plus possible associés à L (la précision) et les objets associés à L dans $\mathcal{D}$ doivent être au maximum inclus dans le support du sous-groupe (le rappel). La F_1-*Mesure* se base sur le rappel et la précision d'un sous-groupe vis à vis du sous-ensemble L à caractériser. Pour un sous-groupe local (d, L), on note : $E10 = |\{o \in \mathcal{O} | o \in supp(d), class(o) \cap L \neq L\}$, $E11 = |\{o \in \mathcal{O} | o \in supp(d), class(o) \cap L = L\}$, et enfin $E01 = |\{o \in \mathcal{O} | o \notin supp(d), class(o) \cap L = L\}$.

On définit alors la précision par $P(d, L) = \frac{E11}{E11+E10}$ et le rappel par $R(d, L) = \frac{E11}{E11+E01}$. La F_1-*Mesure* pour un sous-groupe local (d, L) s'exprime par :

$$F_1(d, L) = \frac{2 \times (P(d, L) \times R(d, L))}{P(d, L) + R(d, L)}. \quad (3)$$

On remarque alors que la F_1-*Mesure* est comprise entre 0 et 1 puisque la précision et le rappel sont compris aussi entre 0 et 1. Plus la valeur de $F_1(d, L)$ est proche de 1 plus le sous-groupe (d, L) caractérise spécifiquement le sous-ensemble de labels L.

Exemple. Afin d'illustrer la méthode ElMM, nous reprenons l'exemple de la Table 2. Soit le sous groupe (d, L) avec $d_1 = \langle MW \leq 151.28, 23 \leq nAT \rangle$ et $L = \{Miel, Vanillé\}$ le sous-ensemble de labels de classe à caractériser, en appliquant la formule de l'équation 3 afin de calculer la mesure de qualité par la F_1-*Mesure* on trouve : $F_1(d, L) = 2/3$ puisque $P(d, L) = 1/2$ et $R(d, L) = 1$.

Objectif d'ElMM. Étant donné $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$, φ, $minSupp$, $maxDesc$, $maxLab$ et k l'objectif de l'approche d'ElMM est de récupérer les k-meilleurs sous-groupes locaux (d, L) au regard de la mesure de qualité $\varphi = F_1(d, L)$ vérifiant les contraintes imposées.

4 Découverte de sous-groupes locaux avec ELMMUT

Dans cette section nous présentons l'algorithme ELMMUT qui répond au problème d'Exceptional local Model Mining (ElMM). Tout d'abord, nous caractérisons l'espace de recherche des sous-groupes. Ensuite, nous décrivons la manière de parcourir cet espace pour produire les sous-groupes en illustrant le pseudo-code de ELMMUT.

Espace de recherche. L'espace de recherche correspond à l'ensemble de tous les sous-groupes locaux, partiellement ordonnés. Un sous-groupe local (d_1, L_1) est une spécialisation d'un second sous-groupe (d_2, L_2), si $d_1 \preceq d_2$ et $L_1 \supseteq L_2$. Ainsi, l'espace de recherche correspond à un treillis dans lequel chaque sous-groupe local est un nœud et le lien entre deux nœuds dénote que le nœud de niveau $i+1$ est une spécialisation du nœud de niveau i par ajout d'une nouvelle classe à caractériser, ou par spécialisation d'une restriction de la description. L'élément le plus

général du treillis correspond au sous-groupe local vide que l'on note $(\langle\rangle, \emptyset)$ en omettant les f_i car aucune restriction n'est effectuée pour tout attribut a_i : $f_i = [min_{Dom(a_i)}, max_{Dom(a_i}]$ si a_i est numérique et $f_i = \emptyset$ si a_i est nominal.

Parcours heuristique de l'espace de recherche. L'algorithme ELMMUT effectue un parcours en profondeur de l'arbre de recherche en partant du plus général (le sous-groupe local vide à la racine de l'arbre) vers le plus spécifique. Le principe algorithmique est donné dans l'Algorithme 1. Pour chaque sous-groupe d'un nœud de l'arbre de recherche, ELMMUT essaie de le spécialiser par une extension de description ou de labels tant que la mesure de qualité est améliorée (fonction *Spécialiser*) (Galbrun et Kimmig, 2014). Cependant, il existe dans le pire des cas $|Dom(C)| + (|\mathcal{A}| \times n(n+1)/2)$ possibilités pour spécialiser un sous-groupe, puisque on peut effectuer jusqu'à $|Dom(C)|$ extensions de labels et $|\mathcal{A}|$ extensions de description pour lesquelles on peut construire $n(n+1)/2$ intervalles possibles (si l'attribut possède n valeurs différentes). Afin de pallier à ce problème d'espace de recherche, nous nous sommes tournés vers une approche heuristique utilisée dans la découverte de sous-groupes et dans la fouille de redescriptions, il s'agit d'une approche de type "beam-search" (recherche par faisceau) (Lowerre, 1976). Cette approche permet d'explorer seulement une partie des branches de l'arbre de recherche : à chaque spécialisation, seulement une partie des possibilités (au maximum $beamWidth$) de spécialisation du sous-groupe va être analysée (cf. ligne 11 de *Spécialiser*).

Élagage par contraintes poussées. Lors de la tentative de spécialisation d'un sous-groupe local (d, L), ELMMUT maintient à jour une liste A_{Cand} des attributs pouvant étendre d et une liste L_{Cand} des labels de classe pouvant être ajoutés à L. L'algorithme parcourt seulement ces deux listes pour essayer de spécialiser (d, L) (ligne 2 à 10 de *Spécialiser*). Si les contraintes ((i), (ii), (iii) et mesure de qualité du sous-groupe spécialisé meilleure que celle de (d, L)) sont respectées pour le sous-groupe spécialisé (d', L'), elles lui sont mises à jour ainsi que A_{Cand} et L_{Cand} : on retire l'élément venant d'être ajouté pour spécialiser (d, L) en (d', L'), et si la contrainte (ii) (respectivement (iii)) est atteinte alors on vide la liste A_{Cand} (respectivement L_{Cand}). L'énumération s'arrête lorsque les deux listes A_{Cand} et L_{Cand} sont vides ou lorsqu'on ne peut pas améliorer la mesure de qualité d'un sous-groupe en le spécialisant.

Optimisation des intervalles à la volée. Pour les attributs numériques, une simple discrétisation en prétraitement n'est pas suffisante. Cette approche est cependant utilisée dans une partie des expérimentations afin de pouvoir se comparer équitablement à l'algorithme de référence pour EMM (DSSD Diverse Subgroup Set Discovery introduit par van Leeuwen et Knobbe (2012)), qui utilise une telle discrétisation. Afin d'obtenir des résultats les meilleurs possibles, le choix des bornes de l'intervalle pour un attribut $a \in \mathcal{A}$ doit se faire à la volée pour tenir compte des spécificités d'un sous-groupe particulier. Le choix des bornes de l'intervalle est alors déterminant. Tester toutes les possibilités d'intervalles n'est pas envisageable car cette méthode peut s'avérer beaucoup trop gourmande en ressources (complexité théorique d'ordre n^2 pour n valeurs différentes). Afin de pallier ce problème, nous avons adopté une méthode de discrétisation proche de celle de Fayyad et Irani (1993), introduite dans la découverte de sous-groupes par Grosskreutz et Rüping (2009). La Figure 1 présente la répartition des valeurs prises par les objets du support d'un sous-groupe local (d, L) pour l'attribut a. Pour optimiser la mesure il faut éliminer du support du sous-groupe un maximum d'objets qui ne sont pas associés au sous-ensemble de labels L à caractériser. Soit $S = (d, L)$ un sous-groupe, et $a \in \mathcal{A}$ un attribut à partir duquel on veut étendre d, on note $\{p_1, \ldots, p_{|a|}\}$ l'ensemble ordonné

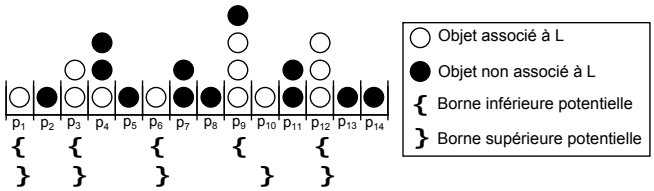

FIG. 1: Processus de discrétisation à la volée.

$(p_1 < p_2 < \cdots < p_{|a|})$ des $|a| \leq |supp(S)|$ valeurs différentes prises par l'ensemble des objets de S pour l'attribut a. On dit alors qu'une valeur p_i des valeurs prises par a est *prometteuse* si le nombre d'objets de $supp(S)$ associés à L possédant la valeur p_i pour a est supérieur ou égal au nombre d'objets de $supp(S)$ non associés à L possédant la valeur p_i. Sinon, on dit qu'elle est *non-prometteuse*. Ainsi, une valeur p_i de a correspond à une borne inférieure potentielle si p_i est *prometteuse* et p_{i-1} est *non-prometteuse*. De plus une valeur p_i de a correspond à une borne supérieure potentielle si p_i est *prometteuse* et p_{i+1} est *non-prometteuse*. Ensuite il suffit de tester tous les intervalles possibles en prenant tous les couples (borne inférieure potentielle, borne supérieure potentielle) et de choisir le meilleur.

5 Expérimentations

5.1 Jeu de données

Nous disposons d'un atlas Arctander (1969), première base de données olfactive établie, qui sert de référence pour les neuroscientifiques. Il met en œuvre 1 689 molécules différentes décrites par 1 704 propriétés physicochimiques numériques (leur volume, leur poids, le nombre d'atomes de carbone qu'elles contiennent, etc...) et sont associées à leur(s) qualité(s) olfactive(s) évaluée(s) par des experts. Les possibles discussions quant à l'obtention de cet atlas, et notamment pour les qualités olfactives, ne sont pas abordées dans ce papier comme il s'agit d'un problème traité par les neuroscientifiques en amont. L'atlas étant clairement multi-labels, on associe chaque molécule à un sous-ensemble de qualités olfactives. En moyenne, chaque molécule est associée à 2.88 qualités olfactives.

A partir de cet atlas, nous avons construit deux jeux de données différents. Dans le premier jeu de données $\mathcal{D}_1$, on ne considère que 43 attributs (propriétés physicochimiques) de l'atlas Arctender, alors que dans le second jeu de données $\mathcal{D}_2$ on en considère 243. La sélection des 43 attributs de $\mathcal{D}_1$ a été faite sur recommandation de l'expert qui assure que ces attributs doivent être déterminants pour la caractérisation de qualités d'odeurs. Les 243 attributs du jeu de données $\mathcal{D}_2$ ont quant à eux été sélectionnés par une analyse de non-corrélation des attributs.

5.2 Résultats quantitatifs

Tout d'abord afin de juger de la performance de l'approche que nous avons mise en place, considérons l'aspect quantitatif des résultats sur le jeu de données d'olfaction. Nous avons exécuté les expérimentations sur une machine avec 8Go de RAM et un processeur cadencé à 3.10GHz. Les résultats quantitatifs obtenus ont été réalisés sur les deux jeux de données

Algorithm 1 ELMMUT.

Entrée : $\mathcal{O}, \mathcal{A}, C, class, \varphi, k, beamWidth, minSupp, maxDescr, maxLab$
Sortie : L'ensemble des sous-groupes locaux R

$R \leftarrow \emptyset$
for all $c \in Dom(C)$ **do**
  $Temp \leftarrow \emptyset$
  Vérifier et Mettre à jour les contraintes pour $(\langle\rangle, \{c\})$
  for all $a \in \mathcal{A}$ **do**
    $f \leftarrow$ Choisir restriction sur a
    Vérifier et Mettre à jour les contraintes pour $(\langle f\rangle, \{c\})$
    Ajouter $(\langle f\rangle, \{c\})$ à $Temp$
  end for
  $Temp \leftarrow$ Conserver les k-meilleurs sous-groupes locaux de $Temp$
  for all $(d, L) \in Temp$ **do**
    Ajouter $Spécialiser(d, L)$ à R
  end for
end for

$Fonction$ **Spécialiser**

Entrée : d, L
Sortie : Le meilleur sous-groupe local (d_{best}, L_{best})

$Temp \leftarrow \emptyset$
for all $c \in$ liste des labels de classe candidats de $(d, L) : L_{Cand}$ **do**
  Vérifier et Mettre à jour les contraintes pour $(d, L \cup \{c\})$
  Ajouter $(d, L \cup \{c\})$ à $Temp$
end for
for all $a \in$ liste des attributs candidats de $(d, L) : A_{Cand}$ **do**
  $f \leftarrow$ Choisir restriction sur a
  Vérifier et Mettre à jour les contraintes pour $(d \cup \{f\}, L)$
  Ajouter $(d \cup \{f\}, L)$ à $Temp$
end for
$Temp \leftarrow$ Conserver les $beamWidth$ meilleurs sous-groupes locaux de $Temp$
for all $(d, L) \in Temp$ **do**
  Mettre à jour (d_{best}, L_{best}) en fonction de $Spécialiser(d, L)$
end for

$\mathcal{D}_1$ et $\mathcal{D}_2$. La Figure 2 présente les différents temps d'exécution de notre approche pour la version de l'algorithme sans la discrétisation à la volée pour les attributs (une discrétisation par effectifs égaux est réalisée a priori pour chaque attribut numérique, comme cela est fait par la méthode EMM). Les trois courbes sont relatives au jeu de données $\mathcal{D}_1$. On remarque que plus on augmente la taille maximale autorisée pour la description ($maxDescr$) ou pour le sous-ensemble de labels à caractériser ($maxLab$), plus le temps d'exécution est long ce qui est tout à fait compréhensible puisque l'algorithme cherche à étendre le plus possible les sous-groupes tant que la mesure de qualité est améliorée. On remarque cependant qu'à partir de $maxDescr = 15$ le temps d'exécution est sensiblement semblable ce qui signifie que même si la taille maximale autorisée augmente les sous-groupes ont une description dont la taille ne va pas au-delà d'un certain seuil : la mesure ne peut plus être améliorée en les étendant. Ce résultat semble être causé à la fois par le paramètre $minSupp$ et par le jeu de données. Effectivement, plus les descriptions sont étendues plus le support a une taille qui tend à diminuer, et puisque l'algorithme est déterministe, avec ce jeu de données, on ne peut excéder une description de taille 15. De même pour la taille maximale du sous-ensemble de labels à caractériser, à partir de 2 ou 3 le temps d'exécution reste le même, ce qui concorde avec le fait qu'en moyenne une

molécule est associée à 2.88 qualités olfactives (au-delà de $n > 3$ qualités olfactives, le nombre de molécules partageant l'ensemble de ces mêmes n qualités olfactives est trop faible et la contrainte du support minimale n'est pas respectée). La Figure 3 présente l'impact du jeu de données et de la discrétisation à la volée via notre technique. Clairement, le nombre d'attributs est un facteur crucial pour l'algorithme ELMMUT, on observe la présence d'un facteur 10 entre le temps d'exécution sur $\mathcal{D}_1$ avec 43 attributs et celui sur $\mathcal{D}_2$ avec 243. L'utilisation de la discrétisation à la volée ne semble pas passer à l'échelle lorsque l'on augmente la taille des descriptions : à partir d'une valeur de 15 pour $maxDescr$ l'exécution dure plus de 12 heures et a donc été avortée. Nous prévoyons des techniques d'optimisation dans le futur.

5.3 Résultats qualitatifs

L'interprétation des résultats est un point central dans le cadre de notre application. Les règles descriptives que nous avons mises en place doivent être capables d'informer et d'aiguiller les neuroscientifiques dans leur recherche. Notre approche, ElMM, en ne caractérisant qu'un sous-ensemble de labels de classe permet alors de correspondre au cas pratique à savoir qu'une molécule ne possède en moyenne que 2.88 qualités olfactives. En observant la Figure 4 qui présente la distribution des qualités au sein du jeu de données entier et d'un sous-groupe obtenu par la méthode EMM, on s'aperçoit clairement que l'interprétation d'un tel résultat est très difficile. On constate des différences entre les distributions du sous-groupe et du jeu de données initial mais cette différence est présente sur beaucoup trop de qualités olfactives à la fois et ainsi l'interprétation d'un tel résultat pour la déduction d'une règle descriptive est infaisable pour un neuroscientifique. La Table 3 présente les 5 meilleurs sous-groupes (du point de vue de la mesure F_1) obtenus après suppression des motifs redondants (on utilise ici la même méthode que Galbrun et Kimmig (2014)). Ces sous-groupes sont issus de la base de données $\mathcal{D}_1$ lorsque la discrétisation à la volée est activée avec $maxDescr = 10$, $maxLab = 2$ et $minSupp = 30$. Seulement un sous-groupe caractérisant plusieurs labels de classe (Floral et Balsamique) est présent, avec une mesure de 0.33 et un support de 38. Sa description contient 9 restrictions. Des sous-groupes ont aussi des descriptions plus courtes. La taille des supports est variable. De plus, dans le jeu de données $\mathcal{D}_2$, lorsque la discrétisation à la volée est désactivée et que $maxDescr = 15$, $maxLab = 3$ et $minSupp = 30$, on obtient 74.6% de sous-groupes dont le sous-ensemble de labels est de taille 1, 22.9% de taille 2 et 2.5% de taille 3.

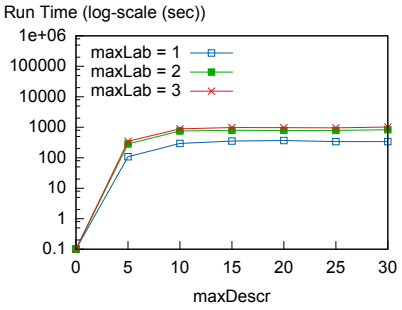

FIG. 2: Temps d'exécution $\mathcal{D}_1$.

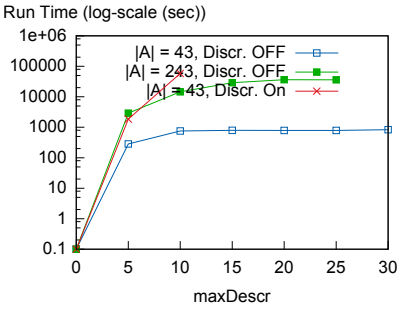

FIG. 3: Temps d'exécution $\mathcal{D}_1$ et $\mathcal{D}_2$.

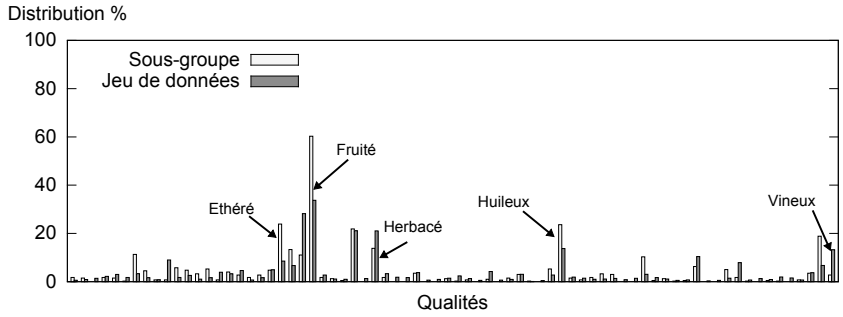

FIG. 4: Distribution des qualités dans un sous-groupe avec EMM.

d	L	$\lvert supp(d)\rvert$	F_1
$\langle 0.116 < X\% < 0.314, 1.0 < nHet < 11.0, 5.159 < Sv < 8.792, 0.0 < nCIC < 0.0, 2.0 < nR03 < 8.0, 0.416 < Ui < 3.551, 4.0 < nArOH < 5.0, 1.0 < nCsp2 < 3.0, 12.0 < nCs < 47.0, 8.0 < nArCOOR < 25.0\rangle$	{Fruité}	654	0.66
$\langle 134.19 < MW < 349.51, 14.0 < nCconj < 100.0, 4.76 < Sv < 8.277, 0.048 < X\% < 0.212, 22.0 < nCs < 49.0, 1.077 < Ui < 3.85, 18.0 < nAB < 49.0\rangle$	{Floral}	740	0.55
$\langle 3.462 < Ui < 3.719, 30.0 < nCconj < 56.0, 40.0 < nAT < 57.0, 35.0 < nO < 50.0\rangle$	{Musqué}	32	0.5
$\langle 2.442 < TPSA(Tot) < 4.028, 4.74 < Sv < 6.095, 2.777 < Ui < 3.921, 0.208 < X\% < 0.31\rangle$	{Huileux}	213	0.44
$\langle 9.0 < nHet < 15.0, 6.095 < Sv < 8.258, 0.0 < Nr05 < 0.0, 2.749 < Ui < 3.517, 25.0 < nAB < 45.0, 2.279 < TPSA(Tot) < 3.334, 24.0 < nRCOOH < 34.0, 21.0 < nCconj < 51.0, 0.074 < X\% < 0.171\rangle$	{Floral, Balsamique}	38	0.33

TAB. 3: Top-5 des sous-groupes locaux.

6 Conclusion

Nous avons présenté la découverte de motifs exceptionnels locaux, une nouvelle méthode de fouille de règles descriptives qui généralise les approches existantes, pour caractériser spécifiquement un sous-ensemble de labels de classe. Nous l'avons appliquée au cas concret de l'olfaction afin de mettre en évidence les liens existant entre les propriétés physicochimiques d'une molécule et ses qualités olfactives. Le pouvoir d'interprétation des résultats et l'information qu'ils véhiculent, permettent d'entrevoir une évolution de la connaissance à propos du phénomène complexe qu'est l'olfaction. De nombreuses expérimentations restent à faire et nous envisageons une exploration interactive inspirée par Galbrun et Miettinen (2012).

Références

Arctander, S. (1969). *Perfume and flavor chemicals :(aroma chemicals)*, Volume 2. Allured Publishing Corporation.

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

Galbrun, E. et A. Kimmig (2014). Finding relational redescriptions. *Machine Learning 96*(3).

Galbrun, E. et P. Miettinen (2012). Siren : an interactive tool for mining and visualizing geospatial redescriptions. In *ACM SIGKDD*, pp. 1544–1547.

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

Kaeppler, K. et F. Mueller (2013). Odor classification : a review of factors influencing perception-based odor arrangements. *Chemical senses 38*(3), 189–209.

Khan, R. M., C.-H. Luk, A. Flinker, A. Aggarwal, H. Lapid, R. Haddad, et N. Sobel (2007). Predicting odor pleasantness from odorant structure : pleasantness as a reflection of the physical world. *The Journal of Neuroscience 27*(37), 10015–10023.

Klösgen, W. (1996). Explora : A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, pp. 249–271. American Association for Artificial Intelligence.

Leman, D., A. Feelders, et A. J. Knobbe (2008). Exceptional model mining. In *ECML/PKDD*, LNCS (5212), pp. 1–16.

Lowerre, B. T. (1976). *The HARPY speech recognition system*. Ph. D. thesis, Carnegie-Mellon Univ., Pittsburgh, PA. Dept. of Computer Science.

Meierhenrich, U. J., J. Golebiowski, X. Fernandez, et D. Cabrol-Bass (2005). De la molécule à l'odeur. *L'actualité chimique* (289), 29.

Novak, P. K., N. Lavrač, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res. 10*, 377–403.

Sezille, C. et M. Bensafi (2013). De la molécule au percept. *Biofutur* (346), 24–26.

van Leeuwen, M. et A. J. Knobbe (2012). Diverse subgroup set discovery. *Data Min. Knowl. Discov. 25*(2), 208–242.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *PKDD*, LNCS (1263).

Summary

Following a complex phenomenon starting with an odorant molecule to the perception in the brain, olfaction is the most difficult sense to understand by neuroscientists. The main challenge is to establish rules on the physicochemical properties of molecules (weight, number of atoms, etc.) to characterize a specific subset of olfactory qualities (fruity, woody, etc.). Subgroup discovery make it possible to find such descriptive rules. However, existing methods provide characterization of either a single label or all the label (exceptional model mining). We then propose an approach for discovering subgroups that characterize only some labels with a new enumeration technique, stemming from redescription mining. We then evaluated this method on an olfactory database provided by the neuroscientists by comparing it with the state-of-the-art algorithm.