Modélisation de concepts par intégrales de Choquet

Grégory Smits[1] Ronald Yager[2] Marie-Jeanne Lesot[3] Olivier Pivert[1]
[1] Université de Rennes - IRISA {gregory.smits,olivier.pivert}@irisa.fr
[2] Machine Intelligence Institute - IONA College yager@panix.com
[3] Sorbonne Université, CNRS, LIP6 marie-jeanne.lesot@lip6.fr

Résumé :
L'utilisation de concepts imprécis, subjectifs et contextuels, comme par ex. celui d'*étudiant prometteur*, permet d'enrichir l'accès aux données, mais la définition de ces concepts est fastidieuse pour un utilisateur. Cet article propose une stratégie, appelée CHOCOLATE, qui requiert uniquement quelques exemples représentatifs du concept, fournis par l'utilisateur, et qui en infère une fonction d'appartenance. CHOCOLATE met en œuvre une intégrale de Choquet pour agréger la pertinence de valeurs observées dans les exemples représentatifs ainsi que la représentativité d'ensembles de ces valeurs. L'approche proposée est en mesure de capturer à la fois les propriétés partagées par la plupart des exemples représentatifs fournis par l'utilisateur et les propriétés spécifiques présentes dans des exemples représentatifs isolés.

Mots-clés :
Concept flou, mesure floue, intégrale de Choquet.

Abstract:
Imprecise and subjective concepts, as e.g. *promising students*, may be used within data mining tasks or database queries to faithfully describe data properties of interest. But the definition of such concepts can be demanding for an end-user. This paper proposes a strategy, called CHOCOLATE, that only requires the user to give a few representative examples of the concept and infers a membership function from them. CHOCOLATE relies on a Choquet integral to aggregate the relevance of individual attribute values observed among all the representative points as well as the representativity of sets of such attribute values. As a consequence, a valuable property of the proposed approach is that it is able to both capture properties shared by most of the user-selected representative data points as well as specific properties possessed by only one specific representative data point.

Keywords:
Fuzzy concept, fuzzy measure, Choquet integral.

1 Introduction

Les bases de données sont le plus souvent décrites par des valeurs précises, numériques ou catégorielles, alors que les utilisateurs, lorsqu'ils expriment des propriétés de ces données, utilisent généralement des concepts imprécis, subjectifs et contextuels. A titre d'exemple, on peut considérer un ensemble d'étudiants décrits par le diplôme qu'ils préparent, leur discipline, le fait qu'ils soient boursiers ou non, leur nombre d'années redoublées et leur moyenne annuelle. Toutes les valeurs associées sont précises. Toutefois, pour décrire les propriétés d'un étudiant, des concepts vagues comme *prometteur*, *dynamique* ou *faible*, dont les définitions sont en effet subjectives et contextuelles, peuvent être utilisés. La définition de tels concepts n'est pas aisée pour un utilisateur, entre autres car des données peuvent les satisfaire pour des raisons différentes. Ainsi un étudiant peut être considéré comme *prometteur* en raison de ses notes ou parce qu'il est en reprise d'études par exemple.

Cet article propose une stratégie nommée CHOCOLATE pour *CHOquet-based COncept LeArning from a Tiny set of Examples*, permettant d'inférer la fonction d'appartenance d'un concept flou à partir d'un ensemble contenant quelques données représentatives de ce concept, fournies par un utilisateur. CHOCOLATE infère une mesure floue quantifiant combien les combinaisons de valeurs d'attributs satisfont un concept sous-jacent, ainsi qu'une mesure quantifiant la pertinence de chaque valeur individuellement. Ces deux mesures sont agrégées par une intégrale de Choquet, qui donne son nom à la méthode, pour définir la fonction d'appartenance du concept.

La section 2 examine la définition de la notion polysémique de *concept* et décrit quelques travaux liés concernant l'apprentissage de concepts. La section 3 présente les principes de l'approche proposée, détaillée ensuite dans la section 4. La section 5 décrit un exemple illustratif d'application de CHOCOLATE à une base de données 2D, soulignant ses caractéristiques.

La section 6 montre l'utilité envisagée de CHOCOLATE dans des applications de requêtes par l'exemple et de systèmes de recommandation.

2 Définition et apprentissage de concepts

Cette section résume quelques travaux concernant la définition et l'apprentissage de concepts, qui varient notamment selon leurs types et les données d'entrée nécessaires.

Extension et intension La notion de concept a été largement discutée dans les domaines de l'intelligence artificielle et du traitement de données, principalement selon la double définition introduite par Wille [11] : un concept a une définition extensionnelle, consistant en la liste de ses membres (données), et une définition intensionnelle (intension en conservant la graphie anglaise), qui correspond à un ensemble de propriétés (valeurs d'attributs), partagées par ses membres. Deux fonctions de dérivation permettent de passer de l'une à l'autre ; l'analyse formelle de concepts (AFC) [4] identifie les concepts comme les points fixes de la combinaison de ces fonctions : un concept est exactement l'ensemble des données qui partagent les propriétés considérées, ces propriétés étant celles qui sont partagées exactement par ces données.

L'approche par prototype [7], inspirée de principes cognitifs [8], constitue un autre exemple : les instances d'un concept (extension), sont utilisées pour identifier les valeurs d'attributs (intension) qui sont partagées par les membres de la classe et qui, de plus, diffèrent des valeurs observées parmi les membres des autres classes.

Cette définition de concept, et son implémentation par l'AFC ou les prototypes, est par nature conjonctive : elle impose que tous les membres de l'extension présentent des valeurs communes d'attributs. Elle ne permet donc en particulier pas de satisfaire un concept pour des raisons différentes, dans une approche disjonctive des concepts. CHOCOLATE relâche cette contrainte.

Concepts flous Certaines approches permettent de définir des concepts flous, considérant des appartenances partielles des données aux concepts. Ainsi, la définition initiale de l'AFC mentionnée ci-dessus a fait l'objet de plusieurs extensions floues [2]. Elles reposent sur l'évaluation du degré de vérité de l'assertion "toutes les données de l'extension possèdent toutes les propriétés de l'intension", où l'extension ou l'intension (ou les deux) peuvent être floues. Elles conservent la définition conjonctive des concepts. L'implémentation de la notion de prototypes [7] repose également sur des sous-ensembles flous. CHOCOLATE considère aussi ce cadre, et fournit une fonction d'appartenance qui associe à chaque donnée un degré d'appartenance au concept.

Concepts discriminants et contre-exemples Un troisième axe concerne les relations entre concepts : en accord avec les principes cognitifs [8], les prototypes flous définissent les concepts par opposition les uns aux autres. Ils prennent en compte des propriétés partagées mais aussi propres aux concepts, c-à-d distinctives. Au contraire, l'AFC traite chaque concept indépendamment des autres.

Cette distinction a des conséquences sur les pré-requis des méthodes d'apprentissage : pour prendre en compte les autres concepts, il est nécessaire de disposer des exemples du concept, mais aussi de contre-exemples, i.e. des exemples des autres concepts. Ce point de vue permet de voir la tâche d'apprentissage de concept comme une tâche de classification, pour distinguer les données qui satisfont un concept de ceux qui ne le satisfont pas. Traiter chaque concept indépendamment signifie que seuls des exemples de membres du concept sont nécessaires.

Cet article considère le cas où les concepts sont appris indépendamment et ne requiert que quelques exemples représentatifs. De plus, pour alléger encore les contraintes imposées à l'utilisateur, CHOCOLATE considère qu'un nombre très faible de tels exemples est disponible. Dans

ce cas, la plupart des méthodes de classification ne peuvent pas être mises en œuvre.

Autres approches liées La tâche de *subspace clustering* [1, 10, 5] peut être considérée comme proche de cette formulation de l'apprentissage de concepts : elle vise à décomposer un ensemble de données en sous-ensembles homogènes et à simultanément déterminer les sous-espaces de l'univers de description dans lesquels ces sous-ensembles sont situés. Aussi, elle fournit une caractérisation des *clusters*, que l'on pourrait interpréter comme des concepts extraits des données. Toutefois, elle ne prend pas en entrée des exemples du concept d'intérêt et conduirait à identifier plusieurs concepts simultanément, en les définissant par opposition les uns par rapport aux autres. De plus elle requiert beaucoup plus de données que le cadre considéré ici. CHOCOLATE peut également être rapprochée de la tâche d'inférence de préférences [3]. Toutefois elle diffère en ce qu'elle ne classe pas les données, mais infère une mesure numérique évaluant à quel point d'autres données sont liées à celles fournies en exemples par l'utilisateur.

3 Principes de CHOCOLATE

3.1 Définitions préliminaires

Notations On note $\mathcal{D}$ l'ensemble contenant n données $\{x_1, x_2, \ldots, x_n\}$ représentées dans l'univers $\mathcal{X}$. Chaque donnée est décrite par les valeurs prises par m attributs $A_1, A_2, \ldots, A_m$. Une donnée x est donc représentée comme $x = \langle x.A_1, x.A_2, \ldots, x.A_m\rangle$, où $x.A_i$ désigne la valeur prise par x dans le domaine de définition D_i de l'attribut A_i. Les attributs peuvent être numériques ou catégoriels. Une *propriété* est un couple constitué d'un attribut et d'une valeur, noté (A_i, p). On dit qu'une donnée x possède la propriété (A_i, p) si $x.A_i = p$ et l'ensemble de propriétés s si $\forall (A_i, p) \in s$, $x.A_i = p$. Réciproquement, les propriétés de x sont tous les couples $(A_i, x.A_i)$ pour $i = 1 \ldots m$.

Définition de concept considérée Par rapport aux axes discutés dans la section 2, les concepts considérés par CHOCOLATE sont flous, ils peuvent avoir une définition disjonctive et sont inférés à partir d'un nombre très faible d'exemples que l'utilisateur considère comme représentatifs, en l'absence de contre-exemples, et qui en constituent une extension partielle.

Formellement, un *concept* C est un sous-ensemble flou de l'univers $\mathcal{X}$, décrit par sa fonction d'appartenance S_c. Il est induit par l'extension partielle fournie par l'utilisateur, notée $\mathcal{E}_C \subseteq \mathcal{D}$ et associée à une étiquette linguistique, le plus souvent un adjectif.

Exemple illustratif Considérons des données correspondant à des étudiants américains décrits par $m = 5$ attributs : *niveau* de domaine D_{niv} = {Bsc.1, Bsc.2, Bsc.3, Mast., PhD}, *majeure* de domaine D_{maj} ={architecture, biologie, informatique, lettres, mathématiques}, *boursier* de domaine {0, 1} indiquant si l'étudiant est boursier ou non, *redoublements* de domaine $[\![1, 10]\!]$ indiquant le nombre d'années redoublées et *note* de domaine $\{A^+, A, A^-, \ldots D^-, F\}$ donnant leur moyenne annuelle.

Considérons de plus que l'utilisateur souhaite définir le concept d'*étudiant prometteur* et fournit l'extension partielle donnée dans le tableau 1. Cet exemple illustre les différences significatives par rapport à la définition de concept [11] : les exemples peuvent n'avoir aucune valeur commune. L'approche proposée est capable de capturer les spécificités de chaque membre de l'extension partielle ainsi que les propriétés partagées par la plupart d'entre eux.

3.2 Vue d'ensemble de CHOCOLATE

CHOCOLATE initie la définition du concept avec les quelques exemples représentatifs fournis par l'utilisateur. Le problème est alors de définir la mesure permettant d'identifier d'autres données satisfaisant le concept d'intérêt, ainsi que le degré auquel elles le satisfont.

Tableau 1 – Exemples représentatifs du concept *étudiant prometteur*.

	niv.	*maj.*	*boursier*	*redoub.*	*note*
x_1	Bsc.1	phys.	1	0	A^+
x_2	PhD	maths	1	0	B^+
x_3	Bsc.3	archi.	0	0	A^+
x_4	Bsc.3	info.	1	0	A^-
x_5	Mast.	biologie	1	0	A^+

Pour une donnée fixée, CHOCOLATE détermine d'abord si ses propriétés individuelles correspondent à celles de l'extension partielle disponible. Nous proposons de considérer que plus une propriété est fréquente parmi les exemples représentatifs, plus son degré individuel de satisfaction est élevé. Cette étape permet d'identifier des propriétés qui sont importantes individuellement pour définir le concept considéré.

De plus, CHOCOLATE identifie des combinaisons de propriétés, qui doivent être à la fois présentes dans la donnée à évaluer et parmi les exemples représentatifs. Cette étape permet d'éviter de combiner des propriétés qui sont individuellement fréquentes mais n'apparaissent jamais ensemble. Elle conduit à une évaluation non additive de l'importance d'un ensemble de valeurs.

Enfin, une intégrale de Choquet est utilisée pour agréger les évaluations individuelles et combinées des propriétés de la donnée considérée : un degré d'appartenance élevé est associé aux données qui présentent un grand ensemble de propriétés individuellement représentatives. De plus, la stratégie d'agrégation proposée permet d'identifier un ensemble de propriétés présentées par un seul exemple représentatif (ou un nombre très réduit de tels exemples), et de donner une importance à ce sous-ensemble même s'il est composé de propriétés qui ne sont pas individuellement présentes fréquemment parmi les exemples représentatifs.

4 Description de CHOCOLATE

Cette section décrit les fonctions proposées pour déterminer l'importance de chaque propriété individuellement et de groupes de propriétés présents dans la donnée à évaluer.

4.1 Fonction δ_i : propriétés individuelles

La fonction qui quantifie l'importance d'une propriété individuelle (A_i, p) par rapport à une extension partielle $\mathcal{E}_C$ est notée δ_i : elle détermine si la valeur p prise par l'attribut A_i est représentative de $\mathcal{E}_C$. Plus p est fréquente, pour l'attribut A_i parmi les éléments de $\mathcal{E}_C$, plus son évaluation est élevée. Ainsi on peut définir

$$\delta_i(p) = \frac{|\{x \in \mathcal{E}_C / x.A_i = p\}|}{|\mathcal{E}_C|}.$$

Pour l'exemple d'extension partielle donnée dans le tableau 1, on obtient par exemple les degrés suivants pour les propriétés (*redoub, 0*) et (*note,* A^+) : $\delta_{\textit{redoub}}(0) = 1$ et $\delta_{\textit{note}}(A^+) = \frac{3}{5}$.

Au lieu d'effectuer une comparaison binaire basée sur une égalité stricte entre la valeur candidate p et les valeurs de l'attribut A_i pour les données de $\mathcal{E}_C$, on peut tenir compte d'une mesure de similarité sim_i définie sur le domaine de l'attribut A_i et d'un seuil η_i en posant

$$\delta_i(p) = \frac{|\{x \in \mathcal{E}_C / sim_i\,(p,\ x.A_i) \geq \eta_i\}|}{|\mathcal{E}_C|}.$$

Pour les valeurs numériques, la mesure de similarité peut par exemple être la valeur absolue de la différence. Elle peut aussi être définie indirectement en spécifiant une partition floue sur le domaine D_i, permettant de prendre en compte l'indistingabilité de certaines valeurs par rapport aux termes flous qui forment la partition (voir par exemple [9]).

4.2 Fonction μ : ensembles de propriétés

Après avoir évalué les propriétés individuellement, des ensembles de telles propriétés sont évalués. Alors qu'une valeur individuelle

est considérée comme importante si elle est fréquente dans l'extension partielle, nous proposons de faire dépendre l'importance d'un ensemble de valeurs de sa taille et de son occurrence (présence/absence) dans l'extension partielle $\mathcal{E}_C$: la mesure μ quantifie à quel point l'ensemble de valeurs correspond à l'un des exemples représentatifs de $\mathcal{E}_C$. Elle est maximale pour un ensemble qui correspond exactement à la description d'un des exemples. En notant $s = \{(A_i, p_i), i = 1 \dots |s|\}$ un tel ensemble de propriétés, on définit donc

$$\mu(s) = \max_{x \in \mathcal{E}_C} \frac{1}{m} |\{(A_i, p_i) \in s / x.A_i = p_i\}|,$$

où m désigne le nombre total d'attributs. On vérifie aisément que μ est une mesure floue : $\mu(\emptyset) = 0$, et si $s \subseteq s'$, alors $\mu(s) \leq \mu(s')$ car, pour tout $x \in \mathcal{E}_C$, $\{(A_i, p_i) \in s / x.A_i = p_i\} \subseteq \{(A_i, p_i) \in s' / x.A_i = p_i\}$.

Si l'on considère $s = \{(\textit{niv.}, \text{PhD}), (\textit{redoub.}, 0), (\textit{note}, B^-)\}$ et $s' = \{(\textit{niv.}, \text{PhD}), (\textit{maj.}, \text{maths}), (\textit{boursier}, 1), (\textit{redoub.}, 2), (\textit{note}, B^+)\}$, et l'extension partielle du tableau 1, on a $\mu(s) = \frac{2}{5}$ (donnée x_2) et $\mu(s') = \frac{4}{5}$ (donnée x_2 aussi) : s' est plus proche d'un exemple représentatif que s qui est seulement observé à un degré 2/5 dans le meilleur des cas.

On peut noter que, comme pour les fonctions δ_i, on peut remplacer dans μ la contrainte d'égalité $x.A_i = p_i$ par une contrainte sur la valeur de similarité $sim_i(p_i,\ x.A_i) \geq \eta_i$.

4.3 Fonction S_C : combinaison

La dernière étape combine les évaluations des propriétés atomiques et des ensembles de propriétés. Nous proposons d'utiliser une intégrale de Choquet pour effectuer cette agrégation des mesures δ_i et μ. Elle permet en effet de tenir compte, dans l'évaluation d'un ensemble de propriétés par rapport à des exemples représentatifs (en utilisant la fonction μ), de ce que ces propriétés sont spécifiques d'un unique exemple ou partagées par de nombreux exemples. Ceci permet de distinguer un ensemble de propriétés présent dans un unique exemple représentatif d'un ensemble de même taille mais partagé par de nombreux exemples.

Les ensembles candidats sont définis comme les ensembles les plus prometteurs d'après leurs valeurs individuelles δ. Notons H_j l'ensemble des j propriétés qui sont le plus en accord avec des exemples représentatifs de $\mathcal{E}_C$, c-à-d qui ont les valeurs de δ les plus grandes. Formellement, soit σ la fonction telle que $\sigma(i) = k$ ssi k est le rang de δ_i classés par ordre décroissant. H_j est alors défini comme $H_j = \{(A_i, x.A_i) / \sigma(i) \leq j\}$ pour $j = 1 \dots m$ et $H_0 = \emptyset$. De plus, soit $\kappa_j(x)$ la $j^{\text{ème}}$ valeur parmi les δ_i, c-à-d le degré de la $j^{\text{ème}}$ propriété la plus représentative de x par rapport au concept C (formellement $\kappa_j(x) = \delta_{\sigma^{-1}(j)}(A_{\sigma^{-1}(j)}, x.A_{\sigma^{-1}(j)})$).

Nous proposons alors de définir le degré d'appartenance d'une donnée x au concept C d'après l'ensemble d'exemples représentatifs $\mathcal{E}_C$ comme

$$\mathcal{S}_C(x) = \sum_{j=1}^{m} (\mu(H_j) - \mu(H_{j-1}))\, \kappa_j(x) \quad (1)$$

On peut noter que l'implication "si $x \in \mathcal{E}_C$, alors $S_C(x) = 1$" n'est vérifiée que si $\mathcal{E}_C = \{x\}$, c-à-d si C est défini par l'unique exemple représentatif x. Ceci vient de ce que la mesure S_C permet de capturer une similarité avec des exemples atypiques de $\mathcal{E}_C$, compris ici comme des exemples ayant une faible ressemblance interne aux autres exemples de C.

Considérons par exemple la donnée $x = \langle \text{PhD},\ \text{littérature},\ 1,\ 0,\ B^- \rangle$ et l'extension partielle d'*étudiant prometteur* du tableau 1. Le calcul de son degré d'appartenance est détaillé dans le tableau 2 qui montre les valeurs individuelles de δ_i, leurs rangs respectifs κ_j, et les valeurs déduites pour H_j et leurs évaluations μ. Le degré d'appartenance de x au concept C est :

$$\mathcal{S}_C(x) = \frac{1}{5} \times 1 + \frac{1}{5} \times \frac{4}{5} + \frac{1}{5} \times \frac{1}{5} + 0 \times 0 + 0 \times 0 = 0.4$$

Ainsi, cet étudiant est plutôt non prometteur, avec un degré légèrement inférieur à 0.5,

Tableau 2 – Calcul de $S_C(x)$ pour $x = \langle \text{PhD},\ \text{littérature},\ 1,\ 0,\ \text{B}^- \rangle$ et $\mathcal{E}_C$ donné dans le tableau 1.

$\delta_{niv.}(\text{PhD}) =$	$\frac{1}{5}$	$= \kappa_3(x)$	$H_1 = \{(\textit{redoub.}, 0)\}$	$\mu(H_1) = \frac{1}{5}$
$\delta_{maj.}(\text{lit.}) =$	0	$= \kappa_4(x)$	$H_2 = H_1 \cup \{(\textit{boursier}, 1)\}$	$\mu(H_2) = \frac{2}{5}$
$\delta_{boursier}(1) =$	$\frac{4}{5}$	$= \kappa_2(x)$	$H_3 = H_2 \cup \{(\textit{niv.}, \text{PhD})\}$	$\mu(H_3) = \frac{3}{5}$
$\delta_{redoub.}(0) =$	1	$= \kappa_1(x)$	$H_4 = H_3 \cup \{(\textit{maj.}, \text{lit.})\}$	$\mu(H_4) = \frac{3}{5}$
$\delta_{note}(B^-) =$	0	$= \kappa_5(x)$	$H_5 = H_4 \cup \{(\textit{note}, \text{B}^-)\}$	$\mu(H_5) = \frac{3}{5}$

d'après sa comparaison aux exemples fournis par l'utilisateur. En effet, il ne partage pas suffisamment de propriétés avec ces exemples, même s'il est boursier et n'a pas redoublé.

5 Etude expérimentale

5.1 Protocole expérimental

Cette section illustre le comportement de CHOCOLATE sur des données jouet en 2 dimensions. CHOCOLATE est appliquée avec les définitions de δ et μ qui considèrent une mesure de similarité et non seulement la stricte égalité, en posant pour sim le complément à 1 de la distance euclidienne normalisée par la valeur maximale observée, et avec $\eta_i = 0.7$. CHOCOLATE est comparée à des approches qui calculent les degrés d'appartenance à partir des distances aux exemples représentatifs, selon deux opérateurs d'agrégation :

$$\mathcal{S}^1_C(x) = \min_{z \in \mathcal{E}_C} d(x, z) \quad (2)$$

$$\mathcal{S}^2_C(x) = \text{avg}_{z \in \mathcal{E}_C} d(x, z) \quad (3)$$

où $d(x, z) \in [0, 1]$ est la distance euclidienne normalisée par la valeur maximale observée.

La figure 1 montre les résultats obtenus par les trois méthodes pour deux extensions partielles $\mathcal{E}_C$ constituées des points représentés par des losanges. Les degrés d'appartenance sont représentés pour tous les points d'une grille régulière par des niveaux de gris. Les points pour lesquels le degré est inférieur à 0.2 ne sont pas montrés.

5.2 Résultats

Cas général Sur le premier graphique, en haut à gauche de la figure 1, un effet non uniforme de CHOCOLATE autour des points de $\mathcal{E}_C$ peut être observé, ainsi qu'un effet de groupe : les degrés d'appartenance sont plus élevés autour des trois exemples représentatifs groupés autour du point (2, 8) qu'à proximité de l'exemple isolé situé en (2, 2) et plus encore que pour le point (8, 6). Ce dernier montre un impact asymétrique sur la définition du concept. Les résultats montrent aussi que CHOCOLATE identifie que $(A_1, 2)$ est une propriété importante : tous les points qui la possèdent ont un degré d'appartenance élevé. La seconde propriété importante est $A_2 \approx 8,5$, dans une moindre mesure. Ceci explique l'absence de symétrie autour de l'exemple représentatif (8, 6).

Ces données jouet montrent clairement les caractéristiques principales de CHOCOLATE : les fonctions δ_i quantifient l'importance de chaque propriété individuellement, et permettent de capturer le fait que les 4 exemples représentatifs de gauche partagent des valeurs de A_1 proches. La mesure floue μ donne ensuite plus d'importance aux points à proximité des trois exemples situés en haut à gauche, parce qu'ils partagent des valeurs similaires sur les deux dimensions. Toutefois, contrairement à une approche uniquement basée sur la distance (voir ci-dessous), la proximité avec au moins un exemple représentatif est également prise en compte, comme le montrent les points de droite.

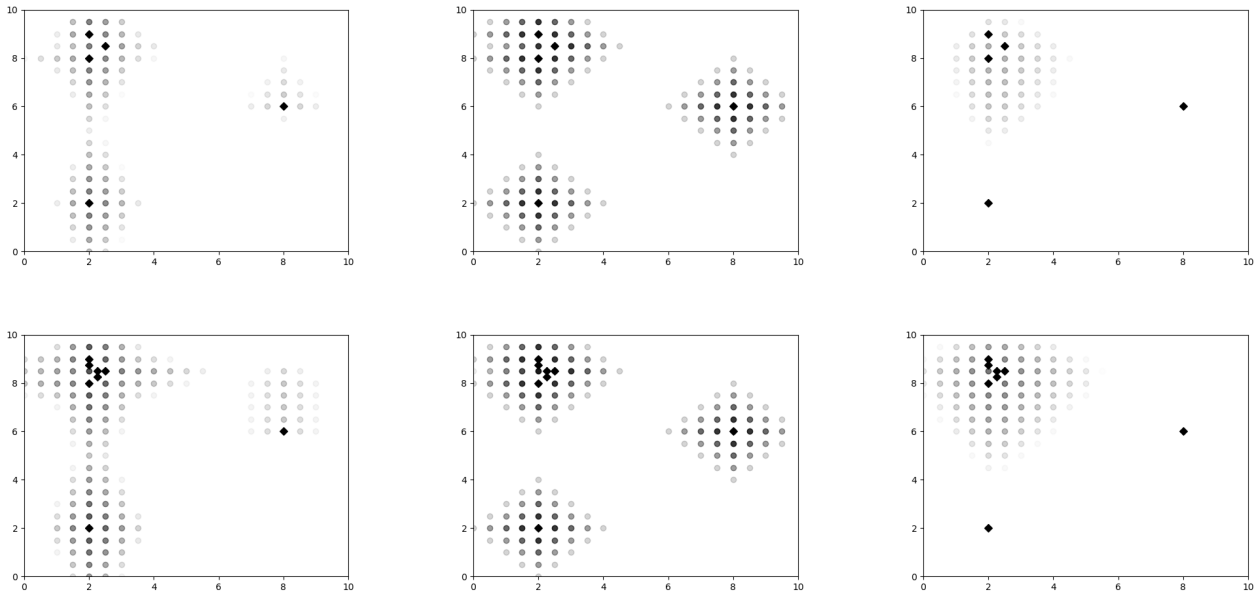

Figure 1 – Concepts flous inférés pour les deux extensions partielles $\mathcal{E}_C$ (une par ligne) constituées des points représentés par des losanges : (gauche) CHOCOLATE, S_C cf Eq. 1, (centre) plus proche voisin S_C^1 cf Eq. 2, (droite) distance moyenne S_C^2 cf Eq. 3.

Cette stratégie offre donc un compromis intéressant entre l'importance de chaque propriété, le nombre de propriétés partagées entre une donnée et un exemple représentatif et le nombre d'exemples représentatifs qui possèdent cette propriété. Ainsi CHOCOLATE est capable d'intégrer une comparaison avec des exemples représentatifs différents, tenant compte de leur diversité, et d'identifier des propriétés partagées par plusieurs de ces exemples.

Sur cette même extension partielle, on peut observer que la méthode par distance minimale aux exemples représentatifs (Eq. 2) se comporte bien comme une méthode par plus proche voisin et conduit à des degrés d'appartenance uniformément distribués autour des exemples fournis. Elle ne permet pas de caractériser le concept sous-jacent. L'approche par distance moyenne (Eq. 3) est très sensible à l'effet de groupe ; d'autres stratégies de normalisation ne permettraient pas une intégration juste des exemples représentatifs isolés.

Cas des concepts disjonctifs La seconde ligne de la figure 1 montre un cas où les exemples représentatifs isolés le sont plus encore, pour illustrer la capacité de CHOCOLATE à donner de l'importance aux propriétés partagées comme aux propriétés rares de l'extension partielle fournie par l'utilisateur : celle-ci est constituée d'un groupe assez compact dans le coin en haut à gauche, et de deux points isolés, l'un étant proche du groupe sur l'axe des abscisses.

Le comportement de CHOCOLATE est intéressant car les fonctions d'appartenance générées sont attirées par le groupe d'exemples représentatifs, mais elles parviennent à tenir compte des exemples isolés et ne les ignorent pas. A nouveau, la présence de deux exemples isolés distants rend les appartenances autour du groupe d'exemples asymétriques, avec un accent mis sur les abscisses proches de 2.

Sur cette extension partielle, les stratégies basées sur les distances ne donnent pas de résultats convaincants : la stratégie par agrégation min ne tient pas compte de la présence du groupe d'exemples pour le traitement des données autour des exemples atypiques. Pour la stratégie par agrégation

moyenne, ce groupe compact donne uniquement du poids à la région dans laquelle ils sont situés.

6 Exemple d'applications

Dans de nombreux contextes applicatifs qui font intervenir une interaction avec l'utilisateur, un problème crucial est de saisir l'intention de celui-ci en requérant le moins possible d'information de sa part. CHOCOLATE peut alors être appliquée pour exploiter au mieux les exemples représentatifs que l'utilisateur fournit.

Ainsi, dans le cadre de l'interrogation de bases de données, les requêtes par l'exemple sont définies par des exemples d'entrée fournis par l'utilisateur ou par son évaluation d'un ensemble d'exemples. Le résultat est un ensemble de tuples qui reflètent ses choix. Plusieurs approches ont été proposées pour inférer des requêtes floues à partir d'exemples, en déterminant des modalités floues qui représentent au mieux les exemples positifs et rejettent les exemples négatifs [6], ou par clustering [12]. Ces approches considèrent néanmoins des requêtes conjonctives. Dans ce contexte, CHOCOLATE permet de construire une requête faisant intervenir des concepts flous complexes qui ne peuvent être exprimés facilement par des combinaisons conjonctives/disjonctives de termes flous atomiques.

Un autre exemple est fourni par le contexte de la recommandation, où un un problème classique est celui du démarrage à froid (*cold start*), pour faire des recommandations à de nouveaux utilisateurs. CHOCOLATE, qui ne nécessite que très peu d'exemples initiaux, peut permettre de réduire la taille de l'historique de l'utilisateur et ouvre ainsi une perspective de recherche.

7 Conclusion et perspectives

La méthode CHOCOLATE proposée construit, à partir d'un petit ensemble de points illustrant différentes significations d'un concept, une fonction d'appartenance floue le représentant. Étant donné le faible nombre d'exemples et l'absence de contre-exemples, la plupart des approches classiques ne peuvent être mises en œuvre. Dans ce contexte, CHOCOLATE permet d'identifier des ensembles de propriétés partagées par les exemples disponibles mais aussi de tenir compte de propriétés spécifiques rares.

Les perspectives de ces travaux incluent des expérimentations au-delà des données jouet illustrant son comportement, en particulier en grandes dimensions ou pour une tâche de recommandation, ainsi que l'effet de la mesure de similarité et du seuil associé. Une autre perspective porte sur une aide à l'utilisateur pour sélectionner les exemples représentatifs, par exemple en envisageant des stratégies issues du domaine des requêtes par l'exemple.

Références

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. of the ACM SIGMOD'98*, pages 94–105, 1998.

[2] R. Bělohlávek and V. Vychodil. What is a fuzzy concept lattice ? In *Proc. of CLA05*, pages 34–45, 2005.

[3] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer, 2010.

[4] B. Ganter and R. Wille. *Formal concept analysis : mathematical foundations*. Springer Science & Business Media, 2012.

[5] M.-J. Lesot. Subspace clustering and some soft variants. In *Proc. of SUM'19*. Springer, 2019.

[6] A. Moreau, O. Pivert, and G. Smits. Fuzzy query by example. In *Proc. of ACM SAC'18*, pages 688–695, 2018.

[7] M. Rifqi. Constructing prototypes from large databases. In *Proc. of IPMU'96*, pages 300–306, 1996.

[8] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and categorization*, pages 27–48. Lawrence Erlbaum associates, 1978.

[9] G. Smits, O. Pivert, and T. N. Duong. On dissimilarity measures at the fuzzy partition level. In *Proc. of IPMU'18*, pages 301–312. Springer, 2018.

[10] R. Vidal. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 28(2) :52–68, 2010.

[11] R. Wille. Restructuring lattice theory : an approach based on hierarchies of concepts. In *Ordered sets*, pages 445–470. Springer, 1982.

[12] S. Zadrozny, J. Kacprzyk, and M. Wysocki. On a novice-user-focused approach to flexible querying : The case of initially unavailable explicit user preferences. In *Proc. of ISDA'10*, pages 696–701, 2010.