Concept Membership Modeling Using a Choquet Integral

Grégory Smits[1], Ronald R. Yager[2], Marie-Jeanne Lesot[3], and Olivier Pivert[1]

[1]University of Rennes – IRISA, UMR 6074, Lannion, France
{gregory.smits,olivier.pivert}@irisa.fr
[2]Machine Intelligence Institute - IONA College, New Rochelle, NY. USA
yager@panix.com
[3]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
marie-jeanne.lesot@lip6.fr

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. However, defining these concepts is a demanding task for the end-user. We thus provide a strategy, called CHOCOLATE, that only requires the user to give a tiny subset of data points that are representative of the concept he/she has in mind, and that infers a membership function from them. This function may then be used to retrieve, from the whole dataset, a ranked list of points that satisfy the concept of interest. CHOCOLATE relies on a Choquet integral to aggregate the relevance of individual attribute values 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

Datasets generally contain points described by precise numerical and categorical values whereas users, when they express properties about data, often use imprecise, complex, context-dependent and subjective concepts. As an illustrative example, consider a set of students described by their marks, prepared diploma, number of repeated years, having a scholarship grant, etc. These values are all precise ones. However, to describe some properties a student may possess, vague concepts as *promising*, *dynamic* or *weak* to name a few, whose definitions are subjective and context-dependent, are more naturally used. The definition of such complex properties is not an easy task for an end-user as data points may satisfy a given concept for very different reasons. A student may for instance be considered *promising* because of his/her marks in a selective major, or for having restarted studies after a long break, among others.

This paper proposes the CHOCOLATE strategy, which stands for CHOquet-based COncept LeArning from a Tiny set of Examples, for inferring the membership function of a fuzzy concept from a small set of representative data points provided by a user. From these examples, CHOCOLATE infers a fuzzy measure that quantifies the extent to which combinations of attribute values match the underlying concept, as well as a measure quantifying the relevance of each attribute value individually. These two quantities are then aggregated using a Choquet integral, that gives the method its name, to determine the degree to which a point satisfies the concept.

The paper is structured as follows. Section **??** discusses the polysemous notion of *concept*, as well as some related works about concept learning. Section **??** presents the principles of the proposed approach, then detailed in Section **??**, that infers a membership function from a small set of representative data points. Section **??** describes an illustrative example on a 2D toy dataset that serves to emphasize the characteristics of the approach. Section **??** shows the usefulness of CHOCOLATE to *fuzzy query by example* and recommendation systems.

2 Related Works on Concept Definition and Learning

This section discusses the notion of a concept, as well as some methods proposed in the literature for learning concepts. Among other things, these methods vary according to the types of concepts they can extract as well as their inputs.

Concept Extent and Intent The notion of a concept has been widely used in artificial intelligence and data management, mainly following the twofold definition introduced by R. Wille [**?**]: a concept has an extensional definition, that consists of the set of its members (i.e. data points), and an intentional one that corresponds to the set of data properties (i.e. attribute values) shared by the members of its extent. It is possible to build, through so-called derivation functions, the extensional set of a concept from its intentional definition, and *vice versa*. Formal Concept Analysis, FCA [**?**], then identifies concepts as fixed points of the combination of these two functions: concepts are the sets of points that are exactly those that possess all the properties in the set, that in turn are shared by exactly the points in the set.

Another example is the fuzzy prototype approach [**?**,**?**] inspired from cognitive definitions of class representatives [**?**]: classes instances, defining the class extent, can be used to identify attribute values that are shared by the class members and that differ from the values observed by the members of other classes, so as to determine the class intent.

This concept definition, and its implementation in FCA and fuzzy prototypes, rely on a conjunctive definition of concepts, that in particular imposes that all members of the extent have common attribute values. It thus does not allow data to satisfy a concept for different reasons, i.e. to have disjunctive concept intents. The proposed CHOCOLATE approach relaxes this constraint, e.g. making it possible to handle the case of *promising student* mentioned in the introduction.

Crisp and Fuzzy Concepts Existing approaches to define and learn concepts can be structured depending on whether they consider crisp or fuzzy definitions, i.e. whether the point assignment to the concepts is binary or weighted.

The initial concept definition [**?**] mentioned above as well as Formal Concept Analysis are binary. They have been extended in several ways to Fuzzy Formal Concept Analysis (e.g. see the comparative study [**?**]), allowing for more flexibility: they rely on assessing the truth degree of the statement "the data in the extent all have the properties of the intent", where both the extent and the intent can be fuzzy sets (or only one of them). They thus still consider a conjunctive definition of concepts. The implementation of prototypes proposed in [**?**,**?**] also relies on fuzzy subsets. The CHOCOLATE approach proposed in this paper also relies on fuzzy concept definition, and outputs a membership function that associates each data point with its membership degree to the concept.

Discriminative Concepts and Counterexamples A third property concerns the relations between concepts. In agreement with cognitive principles [**?**], fuzzy prototypes define a concept in opposition to others. They take into account distinctive and shared features between concepts. On the contrary, FCA deals with each concept independently from the others.

This distinction has consequences on the approach requirements: in order to take other concepts into account, it is necessary to dispose of examples of the considered concept, but also counterexamples, i.e. instances of the other concepts. This point of view opens the way to formalizing the concept learning task as a classification task, distinguishing the data points that satisfy a considered concept from those that do not satisfy it. Dealing with each concept individually means that only representatives of the current concept are required. This paper considers the case where a concept is learned independently from the others, only requiring a few representative points. To alleviate the requirements on the user, it also considers the case where only very few representative examples are available. When the number of available observations is very small, most classification approaches based on the observation of regularities cannot be applied.

Other Related Approaches It may be argued that the considered definition and constraints on concept learning is related to the subspace clustering task [**?**,**?**,**?**]: the latter is an unsupervised task that aims at decomposing a set of points into homogeneous subsets and simultaneously determining the subspaces they are situated in. As a consequence, it provides characterizations of the clusters, that may be interpreted as concepts extracted from the data. However, it does not take as input any example of a concept of interest and it would lead to identify several concepts at once, defining them in opposition to one another. As a consequence, it requires more data than the framework considered in this paper. The proposed CHOCOLATE approach can also be related to preference inference from partial definition [**?**], but it differs by the fact that, instead of ranking the items, it aims at inferring a measure from it to evaluate how much other points are related to the ones provided by the user.

3 Overview of the Proposed CHOCOLATE Approach

After presenting the notations and the toy example used throughout the paper, this section gives an overview of the proposed CHOCOLATE approach and presents its underlying principle, making clear the benefits of using a Choquet integral.

3.1 Notations

The paper relies on the following notations: we consider a dataset $\mathcal{D}$ containing n data points $\{x_1, x_2, \ldots, x_n\}$ from a universe $\mathcal{X}$. Each data point is described by the values it takes on m attributes $A_1, A_2, \ldots, A_m$. A data point x is thus represented as $x = \langle x.A_1, x.A_2, \ldots, x.A_m \rangle$, where $x.A_i$ denotes the value taken by x on the definition domain D_i of attribute A_i. The attributes can be numerical or categorical ones.

Throughout the paper, a *property* is defined as a couple made of an attribute and a value, denoted (A_i, p). A point x is said to possess a property (A_i, p) if $x.A_i = p$ and a set of properties s if $\forall (A_i, p) \in s$, $x.A_i = p$. Reciprocally, the properties of x are all the couples $(A_i, x.A_i)$ for $i = 1 \ldots m$.

As an example throughout the paper, we consider a dataset describing students on $m = 5$ attributes: *level* with domain $D_{level} = \{$Bsc.1, Bsc.2, Bsc.3, Mast., PhD$\}$, *major* with domain $D_{major} = \{$architecture, biology, literature, computer science, maths, physics$\}$, *grant* with domain $\{0, 1\}$ indicating whether the student receives a grant or not, *rep. years* with domain $[\![1, 10]\!]$ indicating the number of repeated years and *mark* with domain $\{A^+, A, A^-, \ldots D^-, F\}$.

3.2 Considered Concept Definition

Based on the discussion from previous section, the concepts considered in this paper are fuzzy ones, they possibly have a disjunctive definition and they are derived from very few instances that the user considers as representative, without disposing of counterexamples. The user-provided representative examples of concept C, denoted by $\mathcal{E}_C$, constitutes a partial extent for C, seen as an initial definition of the extent that has to be completed with appropriate points from the rest of the data with soft assignments.

Definition 1. *A* concept C *is a fuzzy subset of the universe* $\mathcal{X}$*, described by its membership function* $S_c : \mathcal{X} \to [0, 1]$*. It is induced from a user-defined partial extent denoted by* $\mathcal{E}_C \subseteq \mathcal{D}$ *and associated with a linguistic label (generally an adjective from the natural language).*

Example 1. Throughout the paper, we consider that a user wants to define the concept *promising student* and initiates the concept definition by giving the partial extension given in Table **??**.

This example illustrates the significant difference between the definition of a concept in [**?**] and the one used here. Following the discussion given in Section **??**,

Table 1: Partial extension of the concept *promising student*

	level	*major*	*grant*	*rep. years*	*mark*
x_1	Bsc.1	physics	1	0	A^+
x_2	PhD	maths	1	0	B^+
x_3	Bsc.3	architecture	0	0	A^+
x_4	Bsc.3	computer sciences	1	0	A^-
x_5	Mast.	biology	1	0	A^+

we consider that different data points may satisfy the same concept without necessarily sharing all their attribute values. The proposed approach is thus able to capture both the specificities of each member from the partial concept extent as well as properties shared by most of them.

3.3 Underlying Principles of CHOCOLATE

CHOCOLATE initiates the definition of a concept with the few user-given representative points. The central question addressed in this paper is to define a measure to identify other points satisfying the concept of interest, in a fuzzy framework, i.e., quantifying the extent to which they satisfy it.

For a given point, CHOCOLATE first determines if its properties individually match the user-given partial concept extent. We propose to consider that the more frequent a property among the representative elements, the higher its individual matching degree. This step makes it possible to identify properties that appear important individually to define the considered concept.

The complementary step is to identify combinations of properties that are both possessed by the point to be evaluated and by representative elements of the concepts. It aims at preventing from combining properties that are individually frequent but actually are never observed together, thus leading to a non-additive way of evaluating the importance of a set of values.

Finally, the Choquet integral is used to aggregate the individual and combined assessments of the point properties: a high membership degree is assigned to points possessing a large set of properties that are individually representative. In addition, the proposed aggregation strategy makes it possible to identify a specific subset of properties possessed by only one (or a few) of the representative elements of the concept, and to give some importance to this subset even if it is composed of properties that are individually not shared by most of the other representative elements.

4 Proposed Approach to Concept Learning

This section details the proposed functions used in CHOCOLATE to determine the importance of each property individually and of subsets of these properties possessed by a point to evaluate.

4.1 The δ_i Function: Property wrt. Partial Concept Extent

The function that quantifies the importance of a property (A_i, p) wrt. a partial concept extension $\mathcal{E}_C$ is denoted by δ_i: it serves to determine whether a value p taken by attribute A_i is representative of the given partial concept extent. The more p is shared, for attribute A_i, by representative elements of the concept, the more important it is.

As a consequence, the δ_i function checks how often p appears in $\mathcal{E}_C$ for attribute A_i: the degree $\delta_i(p)$ is the highest ($\delta_i(p) = 1$) when p is observed for attribute A_i among all data points in $\mathcal{E}_C$. A first binary matching approach consists in defining:

$$\delta_i(p) = \frac{|\{x \in \mathcal{E}_C / x.A_i = p\}|}{|\mathcal{E}_C|}. \tag{1}$$

Example 2. Using the concept extension given in Table **??**, we obtain the following matching degrees for the properties (*rep. years, 0*) and (*mark,* A^+): $\delta_{rep.\ years}(0) = 1$ and $\delta_{mark}(A^+) = \frac{3}{5}$.

Instead of performing a binary matching, based on strict equality, between the candidate value p and the A_i values of the data points in $\mathcal{E}_C$, a less drastic comparison can be performed allowing some approximation, as

$$\delta_i(p) = \frac{|\{x \in \mathcal{E}_C / sim_i\,(p,\, x.A_i) \geq \eta_i\}|}{|\mathcal{E}_C|}. \tag{2}$$

This relaxed definition for δ_i implies the use of an appropriate similarity measure sim_i on the domain of attribute A_i and the use of a similarity threshold η_i. For numerical attributes, this similarity measure can for instance be the absolute value of the difference, but many other possibilities exist, see e.g. [**?**]. It can also be indirectly defined by specifying a fuzzy partition on the concerned domain D_i, which makes it possible to take into account the indistinguishability of some values wrt. the satisfaction of the fuzzy terms that form the partition, see e.g. [**?**].

4.2 The μ Measure: Set of Properties wrt Partial Concept Extent

Once the properties have been individually evaluated, the importance of *sets* of such properties is quantified. Whereas an individual value is considered important if it frequently appears in the partial concept extent, the importance of a subset of values depends on its size and whether it appears at least once in the partial concept extent. The fuzzy measure μ thus serves to quantify the extent to which the subset of attribute values matches one of the representative data points in $\mathcal{E}_C$. The μ score is maximal if the assessed set of properties exactly corresponds to one of the representative data points in $\mathcal{E}_C$. Denoting $s = \{(A_i, p_i), i = 1 \ldots |s|\}$ such a set of properties, the binary approach defines μ as:

$$\mu(s) = \max_{x \in \mathcal{E}_C} \frac{1}{m} |\{(A_i, p_i) \in s / x.A_i = p_i\}|, \tag{3}$$

where m denotes the number of attributes. It is straightforward to check that μ is a fuzzy measure: $\mu(\emptyset) = 0$, and if $s \subseteq s'$, then $\mu(s) \leq \mu(s')$ as, for any $x \in \mathcal{E}_C$, $\{(A_i, p_i) \in s / x.A_i = p_i\} \subseteq \{(A_i, p_i) \in s' / x.A_i = p_i\}$.

Example 3. Let us consider the two following examples: $s = \{(level,\ \text{PhD}), (rep.\ years, 0), (mark, B^-)\}$ and $s' = \{(level,\ \text{PhD}), (major, \text{maths}), (grant, 1), (rep.\ years, 2), (mark, B^+)\}$, and again the concept extent from Table **??**. Then, $\mu(s) = \frac{2}{5} = 0.4$ (data point x_2) and $\mu(s') = \frac{4}{5} = 0.8$ (data point x_2 again): s' is closer to a representative point than s that is only observed to the level 2/5 in the best case.

As the δ_i's, the μ measure can be turned into a more gradual version by considering a similarity measure instead of a strict equality:

$$\mu(s) = \max_{x \in \mathcal{E}_C} \frac{1}{m} |\{(A_i, p_i) \in s / sim_i(p_i,\ x.A_i) \geq \eta_i|\}. \tag{4}$$

4.3 The S_C Function: Data Point wrt. Partial Concept Extent

The final step combines the evaluations of atomic properties and set of properties. We propose to use the Choquet integral to perform this aggregation of the δ_i and μ evaluations. It especially takes into account, when comparing a set of properties wrt. representative data points (using the μ function), if these evaluated properties are individually specific to one data point or shared by many. This makes it possible to differentiate between a set of properties possessed by only a single representative data point and another set of properties of the same size but shared by several representative data points.

The candidate sets of properties are defined as the set of the most promising ones, according to their individual δ values: let's H_j denote the subset of the j properties that best match the representative data points from $\mathcal{E}_C$, i.e. the j properties with maximal δ values. Formally, let σ be a ranking function such that $\sigma(i) = j$ iff. j is the rank of δ_i in decreasing order. H_j is then defined as $H_j = \{(A_i, x.A_i) / \sigma_i(i) \leq j\}$ for $j = 1 \ldots m$ and $H_0 = \emptyset$. In addition, $\kappa_j(x)$ denotes the j^{th} value among the δ_i, i.e. the matching degree of the j^{th} most representative property possessed by x wrt. concept C (formally $\kappa_j(x) = \delta_{\sigma^{-1}(j)}(x.A_{\sigma^{-1}(j)})$).

We thus propose to define the membership degree of a data point x to concept C based on the set of representative examples $\mathcal{E}_C$ as

$$\mathcal{S}_C(x) = \sum_{j=1}^{m} (\mu(H_j) - \mu(H_{j-1}))\, \kappa_j(x), \tag{5}$$

Let us notice that the implication $x \in \mathcal{E}_C \rightarrow S_C(x) = 1$ holds only if $\mathcal{E}_C = \{x\}$ which means that C is defined by only one prototypical example. This is because the S_c measure makes it possible to capture a similarity with atypical examples in $\mathcal{E}_C$, here understood as an example with possibly low internal resemblance to the other examples from C.

Table 2: Computation details of $S_C(x)$ for $x = \langle$PhD, literature, 1, 0, B$^-\rangle$ and $C =$*promising student* using $\mathcal{E}_C$ from Table **??**.

$\delta_{level}(\text{PhD}) =$	$\frac{1}{4} = \kappa_3(x)$	$H_1 = \{(rep.\ years, 0)\}$	$\mu(H_1) = \frac{1}{5}$
$\delta_{major}(\text{literature}) = 0$	$= \kappa_4(x)$	$H_2 = H_1 \cup \{(grant, 0)\}$	$\mu(H_2) = \frac{2}{5}$
$\delta_{grant}(1) =$	$\frac{3}{4} = \kappa_2(x)$	$H_3 = H_2 \cup \{(level, \text{PhD})\}$	$\mu(H_2) = \frac{3}{5}$
$\delta_{rep.\ years}(0) =$	$1 = \kappa_1(x)$	$H_4 = H_3 \cup \{(major, \text{literature})\}$	$\mu(H_2) = \frac{3}{5}$
$\delta_{mark}(B^-) =$	$0 = \kappa_5(x)$	$H_5 = H_4 \cup \{(mark, B)\}$	$\mu(H_2) = \frac{3}{5}$

Example 4. Consider the data point $x = \langle$PhD, literature, 1, 0, B$^-\rangle$ and the partial extent of the concept *promising student* from Table **??**. Its membership degrees can be computed as detailed in Table **??**: it shows the individual δ_i values, their respective rank κ_j, and the derived H_j together with their μ values. The satisfaction of x wrt. concept C is:

$$\mathcal{S}_C(x) = \frac{1}{5} \times 1 + \frac{1}{5} \times \frac{3}{4} + \frac{1}{5} \times \frac{1}{4} + 0 \times 0 + 0 \times 0 = \frac{2}{5}$$

This means that the considered student is rather not on the side of *promising*, with a membership degree slightly lower than 0.5, based on his/her comparison with the user-provided examples of *promising students*. Indeed, he/she does not share enough set of properties with these examples, even if he/she received a grant and did not repeat years.

5 Empirical Study of CHOCOLATE's Behavior

This section illustrates the behavior of the approach on 2D toy examples. We first consider the example represented on Figure **??**, for which the user provides $\mathcal{E}_C$ of size 4, represented by the five diamond points. CHOCOLATE is applied using the relaxed version of the δ_i and μ functions, respectively defined in Eq. (**??**) and (**??**) setting *sim* to the counter part of the Euclidean distance and with $\eta_i = 0.7$. The inferred membership degrees for all points on a regular grid in the considered universe are shown by the grey levels, the points whose membership degree is lower than 0.2 are not shown.

Results A non-uniform effect around the points in $\mathcal{E}_C$ can be observed, as well as a grouping effect: the membership degrees are higher around the three grouped representative points in the region around (2, 8) than for the isolated point with coordinates (2, 2) and even more so for (8, 6). The latter however still has a non-symmetrical impact on the concept definition. The results also show that CHOCOLATE identifies that $A_1 = 2$ is a very important value: all

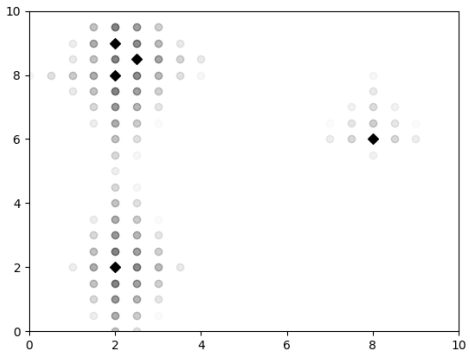

Fig. 1: Membership degrees inferred by CHOCOLATE for the concept with partial extent $\mathcal{E}_C$ defined by the diamond points.

points with this value have a high membership degree. The second important characteristic is $A_2 \approx 8.5$ although to a lesser extent. This explains the absence of symmetry around the representative point with coordinates (8, 6).

This toy dataset clearly shows the main characteristics of our approach: the δ_i functions quantify the importance of each property individually. This allows for instance to capture the fact that the four representative points on the left side of the domain share close A_1 values. Then the fuzzy measure μ gives more importance to the points located close to the three top-left points, because they share close values on the two dimensions. However contrary to a purely distance-based approach (see below), the proximity with at least one of the representatives is also taken into account, this is for instance the case of the points close to the right hand side representative.

This strategy thus enforces an interesting compromise between the importance of each property, the number of properties shared between a data point and a representative, and the number of representative elements of the concepts possessing these properties. This leads to a strategy that is both i) able to handle a comparison with very different representative elements of a concept taking into account their diversity and ii) able to identify properties shared by several of these representative elements.

Comparison to Aggregated Distance Approaches As baseline approaches, we propose to compare CHOCOLATE with membership degrees derived solely from the (Euclidean) distances to the representative points in $\mathcal{E}_C$. Two aggregation operators are considered, defining the membership degrees as:

$$\mathcal{S}_C^1(x) = \min_{z \in \mathcal{E}_C} d(x, z) \qquad \mathcal{S}_C^2 = \text{avg}_{z \in \mathcal{E}_C} d(x, z), \tag{6}$$

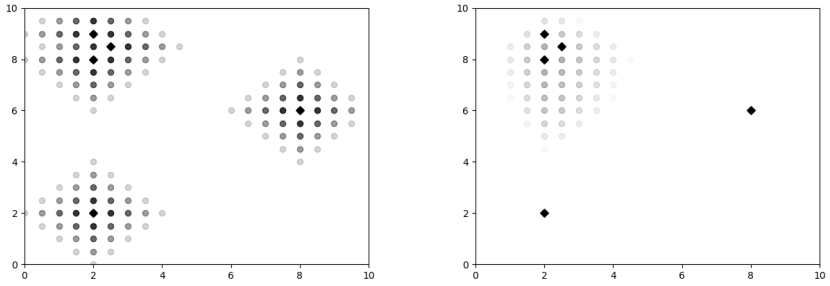

Fig. 2: (Left) Nearest neighbor-based and (Right) Mean-distance-based membership computations.

where $d(x, z) \in [0, 1]$ is the Euclidean distance normalized by the maximum value. The obtained results are shown on Figure **??**: $\mathcal{S}_C^1$ is a nearest-neighbour approach that leads to membership degrees uniformly distributed around the given representative points and does not make it possible to characterize the underlying concept. On the other hand, the mean-distance approach is very sensitive to a grouping effect, even other normalisation strategies would not allow for a fair integration of the isolated representative points.

Case of Disjunctive Concepts As a complement to illustrate CHOCOLATE's ability to both capture and give importance to properties shared by several user-selected representative data points, and to be able to take into account an isolated representative data point, another situation is depicted in Figure **??**, which makes the isolated points more isolated: the user-selected representative data points are composed of a quite compact cluster of data points (in the top left corner) and two isolated data points, one being close to that cluster on the x-axis property.

CHOCOLATE's behavior is interesting as the generated membership function is "attracted" by this cluster of representative data points but still takes into account the isolated representative data points and does not discard them. Again, the presence of two distant representative points, one on the bottom left and one on the right, makes the assignment around the cluster asymmetrical but with an emphasis on the x-values close to 2.

Applied on this same representative data points setting, the minimal and mean distance-based strategies are illustrated in Figure **??**. It shows that for the nearest neighbor strategy the presence of a compact group of representative data points has no impact on the assignment around the atypical data points. As for the mean distance-based approach, this larger compact group of representative data points simply gives more weight to the area they are located in.

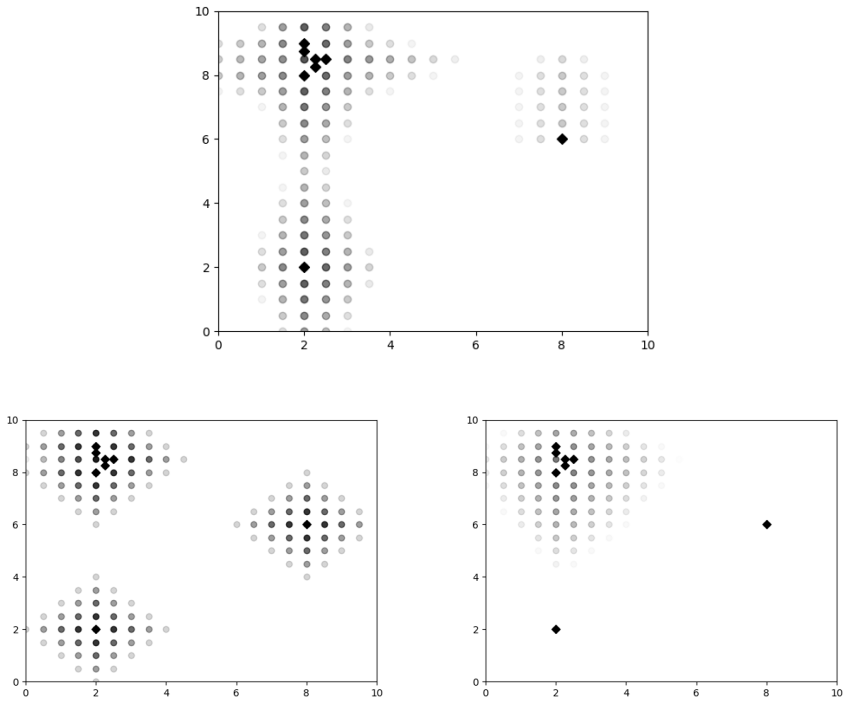

Fig. 3: Membership degrees inferred by CHOCOLATE (Top), nearest-neighbor (Bottom left) and mean-distance (Bottom right), for a partial extent containing a cluster of representative data points and two isolated ones

.

6 Examples of Possible Applications

In many applicative contexts where user interaction is considered, a crucial issue is to capture the user's intent requiring a minimum of knowledge from him/her. In such contexts, the proposed CHOCOLATE approach may be applied with the aim of making the most of only a few examples representing the user's intent. Two examples of such applicative contexts are described here.

Fuzzy Query by Example Query by example is a database and information retrieval paradigm [**?**] that is used to acquire results based either on: one (or several) input tuple(s) provided by the user, or the evaluation by the user of a set of examples (positively, negatively, ...) reflecting the content of the database. The expected output contains elements that are similar to the input tuple(s) provided as example(s), or that reflect the choices of the user if prototypical examples are evaluated.

Table 3: Houses dataset (R: rent, N: nbRooms, L: livingArea, D: distanceToCenter, G: garden)

id	**R**	**N**	**L**	**D**	**G**	$\mathcal{S}(h_i)$	**id**	**R**	**N**	**L**	**D**	**G**	$\mathcal{S}(h_i)$
h_1	450	4	83	0.6	0	0.5	h_6	900	7	99	0.8	1	0.4
h_2	700	5	82	0.5	1	0.3	h_7	450	3	35	0.1	1	0.5
h_3	500	3	45	0.1	1	0.2	h_8	300	2	20	2.1	1	0.1
h_4	600	3	50	1.7	1	0.5	h_9	600	2	50	0.7	1	0.3
h_5	400	5	95	3.5	0	0.3	h_{10}	750	4	50	1.1	0	0.5

In [**?**], we proposed an approach that infers a fuzzy query from user-assessed prototypical examples, based on an algorithm determining the fuzzy modalities (from a fuzzy vocabulary defined on the attribute domains) that best represent the positive examples (and at the same time discards the negative ones). An alternative approach based on clustering is proposed in [**?**]. Both these approaches, however aim to infer simple *conjunctive* fuzzy queries, whereas the technique we propose in the present paper makes it possible to build a fuzzy query involving complex fuzzy concepts that cannot be easily expressed by a conjunctive/disjunctive combination of atomic fuzzy terms.

Example 5. To illustrate how the CHOCOLATE method can be used in a query by example system, let us consider the database describing houses to rent given in Table **??**. A user provides the two following fictional examples describing what he/she considers as *interesting houses*: $h = \langle 400, 3, 40, 0.5, 0\rangle$ and $h' = \langle 800, 6, 99, 0.5, 0\rangle$.

Then, two possibilities are envisaged to retrieve the tuples of interest from the database. If a strict version of the δ_i and μ functions is used (Eq. **??** and **??**), then a query of the following form is executed to return the tuples whose values on the different attributes are present in the partial concept extent:

Q = SELECT * FROM houses
WHERE rent IN (400,800) AND nbRooms IN (3,6) AND ...;

When relaxed versions of the δ_i and μ functions are used (Eq. **??** and **??**), the submitted query returns all the tuples whose values are close enough to those present in the partial concept extent:

Q_r =SELECT * FROM houses
WHERE (sim(rent, 400) <= ETA1 OR sim(rent, 800) <= ETA1) AND ...;

CHOCOLATE is then applied to the tuples returned by the query to rank them according to their satisfaction of the concept partially defined by the user-provided representative tuples. Column $\mathcal{S}(h_i)$ in Table **??** gives the results obtained with the relaxed version of δ_i and μ(Eq. **??** and **??**) and similarity

$$sim_i(v, v') = 1 - \frac{|v - v'|}{\max_h h.A_i - \min_h h.A_i},$$

where A is the considered attribute and the threshold values are 0.7 and 0.9 for ETA1 and ETA2 respectively.

These results show that the membership function built using CHOCOLATE takes into account the specificity of each of the two provided representative

elements: h_6 receives a high score because of its similarity with h', and h_7 because it is rather close to h. It also shows that the distance to the city center and the absence of a garden are important properties that counterbalance other less desired properties, as e.g. in the case of h_{10}.

Data Point Recommendation A second possible applicative context where CHOCOLATE can be useful is that of *recommendation systems*. Most recommendation strategies, especially those based on collaborative filtering, are not able to suggest meaningful recommendations for new users, which is known as the *cold start problem*. A particular interesting aspect of the proposed approach is to be able to find data points that best match the user history, even for a very small available history.

Consider for instance a user who has shown an interest for the manga *Dreamland* by Reno Lemaire, the novel *L'Étranger* by Albert Camus, the science-fiction novel *Ravage* by René Barjavel and the philosophical essay *Discours de la méthode* by René Descartes. Thematically speaking, these books are very different but they all have a French author and half of them are novels. The Choquet integral-based approach would favor French novels by Camus or Barjavel and then French philosophical essays or French comics, etc.

7 Conclusion and Perspectives

When knowledge has to be acquired from user-defined data, one cannot expect to have large labelled training sets. In this work, we face the problem of trying to build, from a small set of data points that illustrate the possible meanings of a concept, a function that can then be used to quantify the extent to which a new data point matches the underlying concept. With such a small set of representative data points and without any negative examples, classical statistical approaches cannot be applied. This is why we propose CHOCOLATE, an alternative strategy based on a generic aggregation framework, namely the Choquet integral involving a fuzzy measure. This method can both identify subsets of properties shared by representative data points of the concept and take into account their specificities, properties possessed by only one or a few representative elements. This approach can be used to recommend data points that best match a user's interest expressed by a few selected examples only.

In this paper, the proposed approach has been applied on toy examples to show its behavior. The interesting obtained results open several perspectives for future works. The first one, despite the topical aspect of the approach, is to assess the relevance of the generated results in a concrete applicative context of a recommendation system, so as to show that relevant recommendations are provided for users having a scarce history. At a more theoretical level, another perspective concerns the extension of CHOCOLATE to handle imprecise descriptions of typical points that form the partial concept extent. Another improvement of the approach would be to help users select the prototypical elements that define

the concept they have in mind. To do so, several strategies from the query by examples field may be envisaged.