On a Plausible Concept-wise Multipreference Semantics and its Relations with Self-organising Maps ⋆

Laura Giordano[1], Valentina Gliozzi[2], and Daniele Theseider Dupré[1]

[1] DISIT - Università del Piemonte Orientale, Alessandria, Italy,
laura.giordano@uniupo.it, dtd@uniupo.it
[2] Center for Logic, Language and Cognition, Dipartimento di Informatica, Università di Torino, Italy, valentina.gliozzi@unito.it

Abstract. In this paper we describe a concept-wise multi-preference semantics for description logic which has its root in the preferential approach for modeling defeasible reasoning in knowledge representation. We argue that this proposal, beside satisfying some desired properties, such as KLM postulates, and avoiding the drowning problem, also defines a plausible notion of semantics. We motivate the plausibility of the concept-wise multi-preference semantics by developing a logical semantics of self-organising maps, which have been proposed as possible candidates to explain the psychological mechanisms underlying category generalisation, in terms of multi-preference interpretations.

1 Introduction

Conditional logics have their roots in philosophical logic. They have been studied first by Lewis [25, 28] to formalize hypothetical and counterfactual reasoning (if A were the case then B) that cannot be captured by classical logic with its material implication. From the 80's they have been considered in computer science and artificial intelligence and they have provided an axiomatic foundation of non-monotonic and common sense reasoning [12, 23]. In particular, preferential approaches [23, 24] to common sense reasoning have been more recently extended to description logics, to deal with inheritance with exceptions in ontologies, allowing for non-strict forms of inclusions, called *typicality or defeasible inclusions* (namely, conditionals), with different preferential semantics [15, 5] and closure constructions [7, 6, 18, 30].

In this paper we consider a "concept-aware" multipreference semantics [13] that has been recently introduced for a lightweight description logic of the $\mathcal{EL}^{\bot}$ family, which takes into account preferences with respect to different concepts, and integrates them into a preferential semantics. To support the plausibility of this semantics we show that it can be used to provide a logical semantics of self-organising maps [22]. Self-organising maps (SOMs) have been proposed as possible candidates to explain the psychological mechanisms underlying category generalisation. They are psychologically

and biologically plausible neural network models that can learn after limited exposure to positive category examples, without any need of contrastive information.

We show that the process of category generalization in self-organising maps produces, as a result, a multipreference model in which a preference relation is associated to each concept (each learned category) and the combination of the preferences into a global one, following the approach in [13], defines a standard KLM preferential model. The model can be exploited to learn or validate conditional knowledge from the empirical data used in the category generalization process, and the evaluation of conditionals can be done by model checking, using the information recorded in the SOM.

Based on the assumption that the abstraction process in the SOM is able to identify the most typical exemplars for a given category, in the semantic representation of a category, we will identify some specific exemplars (namely, the best matching units of the category) as the typical exemplars of the category, thus defining a preference relation among the instances of a category.

The category generalization process can then be regarded as a model building process and, in a way, as a belief revision process. Indeed, initially we have no belief about which is the category of any exemplar. During training, the current state of the SOM corresponds to a model representing the beliefs about the input exemplars considered so far (concerning their category). Each time a new input exemplar is considered, this model is revised adding the exemplar into the proper category.

2 Preliminary: the description logic $\mathcal{EL}^{\bot}$

We consider the description logic $\mathcal{EL}^{\bot}$ of the $\mathcal{EL}$ family [1]. Let N_C be a set of concept names, N_R a set of role names and N_I a set of individual names. The set of $\mathcal{EL}^{\bot}$ *concepts* can be defined as follows: $C \ := A \mid \top \mid \bot \mid C \sqcap C \mid \exists r.C$, where $a \in N_I$, $A \in N_C$ and $r \in N_R$. Observe that union, complement and universal restriction are not $\mathcal{EL}^{\bot}$ constructs. A knowledge base (KB) K is a pair $(\mathcal{T}, \mathcal{A})$, where $\mathcal{T}$ is a TBox and $\mathcal{A}$ is an ABox. The TBox $\mathcal{T}$ is a set of *concept inclusions* (or subsumptions) of the form $C \sqsubseteq D$, where C, D are concepts. The ABox $\mathcal{A}$ is a set of assertions of the form $C(a)$ and $r(a, b)$ where C is a concept, $r \in N_R$, and $a, b \in N_I$.

An *interpretation* for $\mathcal{EL}^{\bot}$ is a pair $I = \langle \Delta, \cdot^I \rangle$ where: Δ is a non-empty domain—a set whose elements are denoted by $x, y, z, \ldots$—and $\cdot^I$ is an extension function that maps each concept name $C \in N_C$ to a set $C^I \subseteq \Delta$, each role name $r \in N_R$ to a binary relation $r^I \subseteq \Delta \times \Delta$, and each individual name $a \in N_I$ to an element $a^I \in \Delta$. It is extended to complex concepts as follows: $\top^I = \Delta$, $\bot^I = \emptyset$, $(C \sqcap D)^I = C^I \cap D^I$ and $(\exists r.C)^I = \{x \in \Delta \mid \exists y.(x, y) \in r^I \text{ and } y \in C^I\}$.
The notions of satisfiability of a KB in an interpretation and of entailment are defined as usual:

Definition 1 (Satisfiability and entailment). *Given an $\mathcal{EL}^{\bot}$ interpretation $I = \langle \Delta, \cdot^I \rangle$:*

- I satisfies an inclusion $C \sqsubseteq D$ if $C^I \subseteq D^I$;

- I satisfies an assertion $C(a)$ if $a^I \in C^I$ and an assertion $r(a, b)$ if $(a^I, b^I) \in r^I$.

Given a KB $K = (\mathcal{T}, \mathcal{A})$, an interpretation I satisfies $\mathcal{T}$ (resp. $\mathcal{A}$) if I satisfies all inclusions in $\mathcal{T}$ (resp. all assertions in $\mathcal{A}$); I is a model *of K if I satisfies $\mathcal{T}$ and $\mathcal{A}$.*

A subsumption $F = C \sqsubseteq D$ (resp., an assertion $C(a)$, $R(a, b)$), is entailed by K, written $K \models F$, if for all models $I = \langle \Delta, \cdot^I \rangle$ of K, I satisfies F.

3 A concept-wise multi-preference semantics

In this section we describe an extension of $\mathcal{EL}^\bot$ with typicality inclusions, defined along the lines of the extension of description logics with typicality [15, 17], but we exploit a different multi-preference semantics [13]. In addition to standard $\mathcal{EL}^\bot$ inclusions $C \sqsubseteq D$ (called *strict* inclusions in the following), the TBox $\mathcal{T}$ will also contain typicality inclusions of the form $\mathbf{T}(C) \sqsubseteq D$, where C and D are $\mathcal{EL}^\bot$ concepts. A typicality inclusion $\mathbf{T}(C) \sqsubseteq D$ means that "typical C's are D's" or "normally C's are D's" and corresponds to a conditional implication $C \mathrel{|\!\sim} D$ in Kraus, Lehmann and Magidor's (KLM) preferential approach [23, 24]. Such inclusions are defeasible, i.e., admit exceptions, while strict inclusions must be satisfied by all domain elements.

Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a set of (arbitrary) $\mathcal{EL}^\bot$ concepts, called *distinguished concepts*. For each concept $C_i \in \mathcal{C}$, we introduce a modular preference relation $<_{C_i}$ which describes the preference among domain elements with respect to C_i. Each preference relation $<_{C_i}$ has the same properties of preference relations in KLM-style ranked interpretations [24], is a modular and well-founded strict partial order, i.e., irreflexive and transitive relation, where: $<_{C_i}$ is *well-founded* if, for all $S \subseteq \Delta$, if $S \neq \emptyset$, then $min_{<_{C_i}}(S) \neq \emptyset$; and $<_{C_i}$ is *modular* if, for all $x, y, z \in \Delta$, if $x <_{C_j} y$ then $x <_{C_j} z$ or $z <_{C_j} y$.

Definition 2 (Multipreference interpretation). *A* multipreference interpretation *is a tuple $\mathcal{M} = \langle \Delta, <_{C_1}, \ldots, <_{C_k}, \cdot^I \rangle$, where:*

(a) Δ is a non-empty domain;
(b) $<_{C_i}$ is an irreflexive, transitive, well-founded and modular relation over Δ;
(d) $\cdot^I$ is an interpretation function, drfined as in $\mathcal{EL}^\bot$ interpretations (see Section 2).

Observe that, given a multipreference interpretation, a triple $\mathcal{M}_{C_i} = \langle \Delta, <_{C_i}, \cdot^I \rangle$, which can be associated to each concept C_i, is a ranked interpretation as those considered for $\mathcal{EL}^\bot$ plus typicality in [19]. The preference relation $<_{C_i}$ allows the set of prototypical C_i-elements to be defined as the set of C_i-elements which are minimal with respect to $<_{C_i}$, i.e., the set $min_{<_{C_i}}(C_i^I)$. As a consequence, the multipreference interpretation above is able to single out the typical C_i-elements, for all distinguished concepts $C_i \in \mathcal{C}$.

The multipreference structures above are at the basis of the semantics for ranked $\mathcal{EL}^\bot$ knowledge bases [13], which have been inspired to Brewka's framework of basic preference descriptions [4]. A *ranked TBox* $\mathcal{T}_{C_i}$ is allowed for each concept $C_i \in \mathcal{C}$, and contains all the defeasible inclusions, $\mathbf{T}(C_i) \sqsubseteq D$, specifying the typical properties of C_i-elements. Ranks (non-negative integers) are assigned to such inclusions; the ones with higher ranks are considered to be more important than the ones with lower ranks.

Consider, for instance, the ranked knowledge base $K = \langle \mathcal{T}_{strict}, \mathcal{T}_{Employee}, \mathcal{T}_{Student}, \mathcal{T}_{PhDStudent}, \mathcal{A} \rangle$, over the set of distinguished concepts $\mathcal{C} = \{Employee, Student, PhDStudent\}$, with empty ABox, and with $\mathcal{T}_{strict}$ the set of strict inclusions:

$Employee \sqsubseteq Adult$ $\quad$ $Adult \sqsubseteq \exists has\_SSN.\top$ $\quad$ $PhdStudent \sqsubseteq Student$

$Young \sqcap NotYoung \sqsubseteq \bot \qquad \exists hasScholarship.\top \sqcap Has\_no\_Scholarship \sqsubseteq \bot$;
where the ranked TBox $\mathcal{T}_{Employee} = \{(d_1, 0), (d_2, 0)\}$ contains the defeasible inclusions:

(d_1) $\mathbf{T}(Employee) \sqsubseteq NotYoung$

(d_2) $\mathbf{T}(Employee) \sqsubseteq \exists has\_boss.Employee$;

the ranked TBox $\mathcal{T}_{Student} = \{(d_3, 0), (d_4, 1), (d_5, 1)\}$ contains the defeasible inclusions:

(d_3) $\mathbf{T}(Student) \sqsubseteq \exists has\_classes.\top$

(d_4) $\mathbf{T}(Student) \sqsubseteq Young$

(d_5) $\mathbf{T}(Student) \sqsubseteq Has\_no\_Scholarship$

and the ranked TBox $\mathcal{T}_{PhDStudent} = \{(d_6, 0), (d_7, 1)\}$ contains the inclusions:

(d_6) $\mathbf{T}(PhDStudent) \sqsubseteq \exists hasScholarship.Amount$

(d_7) $\mathbf{T}(PhDStudent) \sqsubseteq Bright$

Exploiting the fact that for an $\mathcal{EL}^\bot$ knowledge base we can restrict our consideration to finite domains [1], and considering canonical models for $\mathcal{EL}^\bot$ [13] which are large enough to contain a domain element for each possible consistent concept occurring in K (and its complement), the ranked knowledge base K above gives rise to canonical models, where the three preference relations $<_{Employee}$, $<_{Student}$, and $<_{PhDStudent}$ represent the preference among the elements of the domain Δ according to concepts $Employee$, $Student$, and $PhDStudent$, respectively.

While we refer to [13] for the construction of the preference relations $<_{C_i}$'s from a ranked knowledge base K, in the following we will recall the notion of concept-wise multi-preference interpretation which can be obtained by *combining* the preference relations $<_{C_i}$ into a global preference relation $<$. This is needed for reasoning about the typicality of arbitrary $\mathcal{EL}^\bot$ concepts C, which do not belong to the set of distinguished concepts $\mathcal{C}$. For instance, we may want to verify whether typical employed students are young, or whether they have a boss. To answer these query both preference relations $<_{Employee}$ and $<_{Student}$ are relevant, and they might be conflicting for some pair of domain elements as, for instance, tom is more typical than bob as a student ($tom <_{Student} bob$), but more exceptional as an employee ($bob <_{Employee} tom$).

To define a global preference relation, we take into account the specificity relation among concepts, such as, for instance, the fact that a concept like $PhdStudent$ is more specific than concept $Student$. The idea is that, in case of conflicts, the properties of a more specific class (such as that PhD students normally have a scholarship) should override the properties of less specific class (such as that students normally do not have a scholarship).

Definition 3 (Specificity). *A* specificity relation *among concepts in $\mathcal{C}$ is a binary relation $\succ \subseteq \mathcal{C} \times \mathcal{C}$ which is irreflexive and transitive.*

For $C_h, C_j \in \mathcal{C}$, $C_h \succ C_j$ means that C_h is *more specific than* C_j. The simplest notion of *specificity* among concepts with respect to a knowledge base K is based on the subsumption hierarchy: $C_h \succ C_j$ if $\mathcal{T}_{strict} \models_{\mathcal{EL}^\bot} C_h \sqsubseteq C_j$ and $\mathcal{T}_{strict} \not\models_{\mathcal{EL}^\bot} C_j \sqsubseteq C_h$. This is one of the notions of specificity considered for $\mathcal{DL}^N$ [3]. Another one is based on the ranking of concepts in the rational closure of K.

Let us recall the notion of concept-wise multipreference interpretation [13].

Definition 4 (concept-wise multipreference interpretation). *A* concept-wise multipreference interpretation *(or cw^m-interpretation) is a tuple* $\mathcal{M} = \langle \Delta, <_{C_1}, \ldots, <_{C_k}, <, \cdot^I \rangle$ *such that:*

(a) Δ *is a non-empty domain;*
(b) for each $i = 1, \ldots, k$, $<_{C_i}$ *is an irreflexive, transitive, well-founded and modular relation over* Δ*;*
(c) $<$ *is a (global) preference relation over* Δ *defined from* $<_{C_1}, \ldots, <_{C_k}$ *as follows:*

$$x < y \text{ iff } \begin{array}{l} (i)\ x <_{C_i} y, \text{ for some } C_i \in \mathcal{C}, \text{ and} \\ (ii)\ \text{for all } C_j \in \mathcal{C},\ x \leq_{C_j} y \text{ or } \exists C_h (C_h \succ C_j \text{ and } x <_{C_h} y) \end{array}$$

(d) $\cdot^I$ *is an interpretation function, as defined for* $\mathcal{EL}^\bot$ *interpretations (see Section 2), with the addition that, for typicality concepts, we let:*

$$(\mathbf{T}(C))^I = min_<(C^I)$$

where $Min_<(S) = \{u : u \in S \text{ and } \nexists z \in S \text{ s.t. } z < u\}$.

Relation $<$ is defined from $<_{C_1}, \ldots, <_{C_k}$ based on a *modified* Pareto condition: $x < y$ holds if there is at least a $C_i \in \mathcal{C}$ such that $x <_{C_i} y$ and, for all $C_j \in \mathcal{C}$, either $x \leq_{C_j} y$ holds or, in case it does not, there is some C_h more specific than C_j such that $x <_{C_h} y$ (preference $<_{C_h}$ in this case overrides $<_{C_j}$). The idea is that, for two PhD students (who are also students) Bob and Mary, if $mary <_{Student} bob$ and $bob <_{PhDStudent} mary$, we will have $bob < mary$, that is, Bob is regarded as being globally more typical than Mary as he satisfies more properties of typical PhD students wrt Mary although Mary may satisfy additional properties of typical students wrt Bob.

It has been proven [13] that, given a cw^m-interpretation $\mathcal{M} = \langle \Delta, <_{C_1}, \ldots, <_{C_k}, <, \cdot^I \rangle$, the relation $<$ is an irreflexive, transitive and well-founded relation. Hence, the triple $\mathcal{M}' = \langle \Delta, <, \cdot^I \rangle$ is a KLM-style preferential interpretation, as those introduced for $\mathcal{EL}^\bot$ with typicality [16] (and it is not necessarily a modular interpretation). A cw^m-model of a ranked $\mathcal{EL}^\bot$ knowledge base K is then defined as a specific preferential interpretation which builds over the preference relations $<_{C_i}$, constructed from the ranked TBoxes $\mathcal{T}_{C_i}$, and satisfying all strict inclusions and assertions in K. The notion of cw^m-entailment, defined in the obvious way, satisfies the KLM postulates of a preferential consequence relation, and does not suffer from the drowning problem, a well known problem of the rational closure and System Z [29, 2], roughly speaking the problem that, if a subclass of C is exceptional for a given aspect, it is exceptional tout court and does not inherit any of the typical properties of C. We refer to [13] for a discussion on the properties of entailment through some example. In the next section we motivate the plausibility of this concept-wise multipreference semantics showing that it is well suited to provide a semantic characterization of self-organising maps [22].

4 Self-organising maps

Self-organising maps (SOMs, introduced by Kohonen [22]) are particularly plausible neural network models that learn in a human-like manner. In particular: SOMs learn to

organize stimuli into categories in an *unsupervised* way, without the need of a teacher providing a feedback; they can learn with just a few positive stimuli, without the need for negative examples or contrastive information; they reflect basic constraints of a plausible brain implementation in different areas of the cortex [27], and are therefore biologically plausible models of category formation; have proven to be capable of explaining experimental results.

In this section we shortly describe the architecture of SOMs and report Gliozzi and Plunkett's similarity-based account of category generalization based on SOMs [20]. In brief, in [20] the authors judge a new stimulus as belonging to a category by comparing the distance of the stimulus from the category representation to the precision of the category representation.

SOMs consist of a set of neurons, or units, spatially organized in a grid [22], as in Figure 1. Each map unit u is associated with a weight vector w_u of the same dimension-

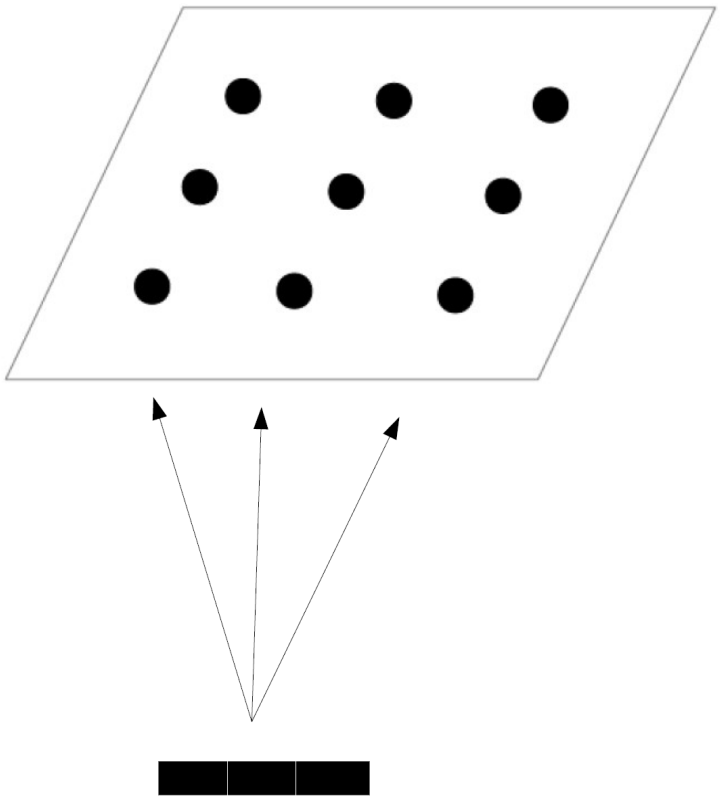

Fig. 1. An example of SOM. The set of rectangles stands for the input presented to the SOM (in the example the input is three-dimensional). This is presented to *all* neurons of the SOM (these are the neurons in the upper grid) in order to find the BMU.

ality as the input vectors. At the beginning of training, all weight vectors are initialized to random values, outside the range of values of the input stimuli. During training, the input elements are sequentially presented to all neurons of the map. After each presentation of an input x, the *best-matching unit* (BMU_x) is selected: this is the unit i whose weight vector w_i is closest to the stimulus x (i.e. $i = \arg\min_j \|x - w_j\|$).

The weights of the best matching unit and of its surrounding units are updated in order to maximize the chances that the same unit (or its surrounding units) will be selected as the best matching unit for the same stimulus or for similar stimuli on subsequent presentations. In particular, it reduces the distance between the best matching unit's weights (and its surrounding neurons' weights) and the incoming input. Furthermore, it organizes the map topologically so that the weights of close-by neurons are updated in a similar direction, and come to react to similar inputs. We refer to [22] for a complete description.

The learning process is incremental: after the presentation of each input, the map's representation of the input (and in particular the representation of its best-matching unit) is updated in order to take into account the new incoming stimulus. At the end of the whole process, the SOM has learned to organize the stimuli in a topologically significant way: similar inputs (with respect to Euclidean distance) are mapped to close by areas in the map, whereas inputs which are far apart from each other are mapped to distant areas of the map.

Once the SOM has learned to categorize, to assess category generalization, Gliozzi and Plunkett [20] define the map's disposition to consider a new stimulus y as a member of a known category C as a function of the *distance* of y from the *map's representation* of C. They take a minimalist notion of what is the map's category representation: this is the ensemble of best-matching units corresponding to the known instances of the category. They use BMU_C to refer to the map's representation of category C and define category generalization as depending on two elements:

- the distance of the new stimulus y with respect to the category representation
- *compared to* the maximal distance from that representation of all known instances of the category

This captured by the following notion of *relative distance* (*rd* for short) [20] :

$$rd(y, C) = \frac{min\|y - BMU_C\|}{max_{x \in C}\|x - BMU_x\|} \tag{1}$$

where $min\|y - BMU_C\|$ is the (minimal) Euclidean distance between y and C's category representation, and $max_{x \in C}\|x - BMU_x\|$ expresses the *precision* of category representation, and is the (maximal) Euclidean distance between any known member of the category and the category representation.

With this definition, a given Euclidean distance from y to $C's$ category representation will give rise to a higher *relative distance rd* if the maximal distance between C and its known examples is low (and category representation is precise) than if it is high (and category representation is coarse). As a function of the relative distance above, Gliozzi and Plunkett then define the *map's Generalization Degree* of category C membership to a new stimulus y.

It was observed that the above notion of relative distance (Equation 1) requires there to be a memory of some of the known instances of the category being used (this is needed to calculate the denominator in the equation). This gives rise to a sort of hybrid model in which category representation and some exemplars coexist. An alternative way of formulating the same notion of relative distance would be to calculate *online* the distance between known category instance currently examined and the representation of the category being formed.

By judging a new stimulus as belonging to a category by comparing the distance of the stimulus from the category representation to the precision of the category representation, Gliozzi and Plunkett demonstrate [20] that the Numerosity and Variability effects of category generalization, described by Griffiths and Tenenbaum [32], and usually explained with Bayesian tools, can be accommodated within a simple and psychologically plausible similarity-based account, which contrasts what was previously maintained. In the next section, we show that their notion of relative distance can also be used as a basis for a logical semantics for SOMs.

5 Relating self-organising Maps and multi-preference models

We aim at showing that, once the SOM has learned to categorize, we can regard the result of the categorization as a multipreference interpretation. Let X be the set of input stimuli from different categories, $C_1, \ldots, C_k$, which have been considered during the learning process.

For each category C_i, we let BMU_{C_i} be the ensemble of best-matching units corresponding to the input stimuli of category C_i, i.e., $BMU_{C_i} = \{BMU_x \mid x \in X \text{ and } x \in C_i\}$. We regard the learned categories $C_1, \ldots, C_k$ as being the concept names (atomic concepts) in the description logic and we let them constitute our set of distinguished concepts $\mathcal{C} = \{C_1, \ldots, C_k\}$.

To construct a multi-preference interpretation we proceed as follows: first, we fix the *domain* Δ^s to be the space of all possible stimuli; then, for each category (concept) C_i, we define a preference relation $<_{C_i}$, exploiting the notion of relative distance of a stimulus y from the map's representation of C_i. Finally, we define the interpretation of concepts.

Let Δ^s be the set of all the possible stimuli, including all input stimuli ($X \subseteq \Delta^s$) as well as the best matching units of input stimuli (i.e., $\{BMU_x \mid x \in X\} \subseteq \Delta^s$). For simplicity, we will assume that the space of input stimuli is finite.

Once the SOM has learned to categorize, the notion of relative distance $rd(x, C_i)$ of a stimulus x from a category C_i introduced above can be used to build a binary preference relation $<_{C_i}$ among the stimuli in Δ^s w.r.t. category C_i as follows: for all $x, x' \in \Delta^s$,

$$x <_{C_i} x' \text{ iff } rd(x, C_i) < rd(x', C_i) \tag{2}$$

Each preference relation $<_{C_i}$ is a strict partial order relation on Δ^s. The relation $<_{C_i}$ is also well-founded as we have assumed Δ^s to be finite.

We exploit this notion of preference to define a multipreference interpretation associated with the SOM, and than a cw^m-model of the SOM. In the following we restrict

the DL language to the fragment of $\mathcal{EL}^{\bot}$ (plus typicality) not admitting roles, as in the self-organising map we do not have a representation of role names.

Definition 5 (Multipreference-model of a SOM). *The* multipreference-model of the SOM *is a multipreference interpretation* $\mathcal{M}^s = \langle \Delta^s, <_{C_1}, \ldots, <_{C_k}, \cdot^I \rangle$ *such that:*

(i) Δ^s *is the set of all the possible stimuli, as introduced above;*
(ii) for each $C_i \in \mathcal{C}$, $<_{C_i}$ *is the preference relation defined by equivalence (2).*
(iii) the interpretation function $\cdot^I$ *is defined for concept names (i.e. categories)* C_i *as follows:*

$$C_i^I = \{y \in \Delta^s \mid rd(y, C_i) \leq rd_{max,C_i}\}$$

where rd_{max,C_i} *is the maximal relative distance of an input stimulus* $x \in C_i$ *from category* C_i*, that is,* $rd_{max,C_i} = max_{x \in C_i}\{rd(x, C_i)\}$*. The interpretation function* $\cdot^I$ *is extended to complex concepts according to Definition 2.*

Informally, we interpret as C_i-elements those stimuli whose relative distance from category C_i is not larger than the relative distance of any input exemplar belonging to category C_i. Given $<_{C_i}$, we can identify the most typical C_i-elements wrt $<_{C_i}$ as the C_i-elements whose relative distance from category C_i is minimal, i.e., the elements in $min_{<_{C_i}}(C_i^I)$. Observe that the best matching unit BMU_x of an input stimulus $x \in C_i$ is an element of Δ^s. Hence, for $y = BMU_x$, the relative distance $rd(y, C_i)$ of y from category C_i is 0, as $min \mid\mid y - BMU_{C_i} \mid\mid = 0$. Therefore, $min_{<_{C_i}}(C_i^I) = \{y \in \Delta^s \mid rd(y, C_i) = 0\}$ and $BMU_{C_i} \subseteq min_{<_{C_i}}(C_i^I)$.

5.1 Evaluation of concept inclusions by model checking

We have defined a multipreference interpretation $\mathcal{M}^s$ where, in the domain Δ^s of the possible stimuli, we are able to identify, for each category C_i, the C_i-elements as well as the most typical C_i-elements wrt $<_{C_i}$. We can exploit $\mathcal{M}^s$ to verify which inclusions are satisfied by the SOM by *model checking*, i.e., by checking the satisfiability of inclusions over model $\mathcal{M}^s$. This can be done both for strict concept inclusions of the form $C_i \sqsubseteq C_j$ and for defeasible inclusions of the form $\mathbf{T}(C_i) \sqsubseteq C_j$, where C_i and C_j are concept names (i.e., categories).

For the verification that a typicality inclusion $\mathbf{T}(C_i) \sqsubseteq C_j$ is satisfied in $\mathcal{M}^s$ we have to check that the most typical C_i elements wrt $<_{C_i}$ are C_j elements, that is $min_{<_{C_i}}(C_i^I) \subseteq C_j^I$. Note that, besides the elements in BMU_{C_i}, $min_{<_{C_i}}(C_i^I)$ may contain other elements of Δ^s having relative distance 0 from C_i. As we do not know, for all the possible input stimuli in Δ^s, whether they belong to $min_{<_{C_i}}(C_i^I)$ or to C_j^I, as an approximation, we only check that all elements in BMU_{C_i} are C_j elements, that is:

$$\text{for all input stimuli } x \in C_i, rd(BMU_x, C_j) \leq rd_{max,C_j} \tag{3}$$

Let the relative distance of BMC_{C_i} from C_j be defined as

$$rd(BMC_{C_i}, C_j) = max_{x \in C_i}\{rd(BMU_x, C_j)\}$$

i.e., as the maximal relative distance of any BMU_x, for $x \in C_i$, and C_j. Then we can rewrite condition (3) simply as

$$rd(BMC_{C_i}, C_j) \leq rd_{max,C_j}.$$

Observe that the relative distance $rd(BMC_{C_i}, C_j)$ also gives a measure of plausibility of the defeasible inclusion $\mathbf{T}(C_i) \sqsubseteq C_j$: the lower is the relative distance of BMU_{C_i} from C_j, the more plausible is the defeasible inclusion $\mathbf{T}(C_i) \sqsubseteq C_j$.

Verifying that a strict inclusion $C_i \sqsubseteq C_j$ is satisfied, requires to check that C_i^I is included in C_j^I. Exploiting the fact that the map is organized topologically, and using the relative distance $rd(BMC_{C_i}, C_j)$ of BMC_{C_i} from C_j, we verify that the relative distance of BMC_{C_i} from C_j plus the maximal relative distance of a C_i-element from C_i is not greater than the maximal relative distance of a C_j-element from C_j:

$$rd(BMC_{C_i}, C_j) + rd_{max,C_i} \leq rd_{max,C_j} \tag{4}$$

where $rd_{max,C} = max_{y \in C}\{rd(y, C)\}$. That is, the C_i-element most distant from C_j is nearer to C_j than the most distant C_j-element.

Computing conditions (3) and (4) on the SOM, may be non trivial, depending on the number of input stimuli that have been considered in the learning phase (the size of the set X of input exemplars). However, from a logical point of view, this is just model checking. Gliozzi and Plunkett have considered self-organising maps that are able to learn from a limited number of input stimuli, although this is not generally true for all self-organising maps [20].

5.2 Combining preferences into a preferential interpretation

The multipreference interpretation $\mathcal{M}^s$ introduced in Definition 5 allows to determine the set of C_i-elements for all learned categories C_i and to define the most typical C_i-elements, exploiting the preference relation $<_{C_i}$. However, we are not able to define the most typical $C_i \sqcap C_j$-elements just using a single preference. Starting from $\mathcal{M}^s$, we construct a concept-wise multipreference interpretation $\mathcal{M}^{som}$ that combines the preferential relations in $\mathcal{M}^s$ into a global preference relation $<$, and provides an intepretation to all typicality concepts as, for instance, $\mathbf{T}(C_i \sqcap C_j \sqcap C_h)$. The interpretation $\mathcal{M}^{som}$ is constructed from $\mathcal{M}^s$ according to Definition 4.

The construction exploits a notion of specificity. Observe that the specificity relation between two concepts C_i and C_j can be determined based on the single model $\mathcal{M}^s$ of the SOM. $C_i \succ C_j$ if $C_i \sqsubseteq C_j$ is satisfied in $\mathcal{M}^s$ and $C_j \sqsubseteq C_i$ is not satisfied in $\mathcal{M}^s$.

Definition 6 (cw^m-model of a SOM). *The* cw^m-model of a SOM *is a cw^m-interpretation $\mathcal{M}^{som} = \langle \Delta^s, <_{C_1}, \ldots, <_{C_k}, <, \cdot^I \rangle$, such that the tuple $\langle \Delta^s, <_{C_1}, \ldots, <_{C_k}, \cdot^I \rangle$ is a multipreference model of the SOM according to Definition 5, and $<$ is the global preference relation defined from $<_{C_1}, \ldots, <_{C_k}$, according as in Definition 4, point (c).*

In particular, in $\mathcal{M}^{som}$, as in all cw^m-interpretations (see Definition 4), the interpretation of typicality concepts $\mathbf{T}(C)$ is defined based on the global preference relation $<$ as $(\mathbf{T}(C))^I = min_<(C^I)$, for all concepts C. Here, we are considering concepts in

the fragment of $\mathcal{EL}^{\bot}$ language without roles, which are built from the concept names $C_1, \ldots, C_n$ (the learned categories). The model $\mathcal{M}^{som}$ can be considered a sort of (unique) canonical model for the SOM, representing what holds in that state of the SOM (e.g., after the learning phase). The logical inclusions that "follow from the SOM" are therefore the inclusions that hold in the single model $\mathcal{M}^{som}$. The situation is similar to the case of Horn clauses, where there is a unique minimal canonical model describing all the (atomic) logical consequences of the knowledge base.

As $\mathcal{M}^{som}$ is a cw^m-interpretation, the triple $\langle \Delta^s, <, \cdot^I \rangle$ is a KLM style preferential interpretation [23, 24]. It follows that the model $\mathcal{M}^{som}$ provides a logical semantics for the SOM which is well-defined, as $\mathcal{M}^{som}$ determines a preferential interpretation and, also, a preferential consequence relation, satisfying all KLM properties of a preferential consequence relations.

The verification of arbitrary defeasible inclusions on $\mathcal{M}^{som}$ can, in principle, be done by model checking, but it might require considering all the possibly many input stimuli, i.e., all domain elements in Δ^s, which may be unfeasible in practice. As an alternative, the identification of the set of strict and defeasible inclusions satisfied by the SOM over the learned categories $C_1, \ldots, C_k$ (as done in Section 5.1), allows to define an $\mathcal{EL}^{\bot}$ knowledge base K and to reason on it symbolically, using for instance an approach similar to the one described in Section 3 for ranked knowledge bases. In particular, Answer Set Programming, and *asprin*, have been used to achieve defeasible reasoning under the multipreference approach for the lightweight description logic $\mathcal{EL}^{+}_{\bot}$ [13]. Ranked knowledge bases have been considered, where the rank of defeasible inclusions provides a measure of plausibility of the defeasible inclusion, and multipreference entailment has been reformulated as a problem of computing preferred answer sets. As we have seen, a measure of plausibility can as well be assigned to the defeasible inclusions satisfied by the SOM.

5.3 Category generalization process as iterated belief revision

We have seen that one can give an interpretation of a self-organising map after the learning phase, as a preferential model. However, the state of the SOM during the learning phase can as well be represented as a multipreference model (precisely in the same way). During training, the current state of the SOM corresponds to a model representing the beliefs about the input stimuli considered so far (beliefs concerning the category of the stimuli).

The category generalization process can then be regarded as a model building process and, in a way, as a belief revision process. Initially we do not know the category of the stimuli in the domain Δ^s. In the initial model, call it $\mathcal{M}_0^{som}$ (over the domain Δ^s) the interpretation of each concept C_i is empty. $\mathcal{M}_0^{som}$ is the model of a knowledge base K_0 containing a strict inclusion $C_i \sqsubseteq \bot$, for all C_i.

Each time a new input stimulus ($x \in C_i$) is considered, the model is revised adding the stimulus x (and its best matching unit BMU_x) into the proper category (C_i). Not only the category interpretation is revised by the addition of x and BMU_x in C_i^I (so that $C_i \sqsubseteq \bot$ does not hold any more), but also the associated preference relation $<_{C_i}$ is revised as the addition of BMU_x modifies the set of best matching units BMU_{C_i} for

category C_i, as well as the relative distance $rd(y, C_i)$ of a stimulus y from C_i. That is, a revision step may change the set of conditionals which are satisfied by the model.

At the end of the training process, the final state of the SOM is captured by the model $\mathcal{M}^{som}$ obtained by a sequence of revision steps which, starting from $\mathcal{M}_0^{som}$, gives rise to a sequence of models $\mathcal{M}_0^{som}$,$\mathcal{M}_{i_1}^{som}, \ldots, \mathcal{M}_{i_r}^{som}$ (with $\mathcal{M}^{som} = \mathcal{M}_{i_r}^{som}$). At each step the knowledge base is not represented explicitly, but the model $\mathcal{M}_{i_j}^{som}$ of the knowledge base at step j is used to determine the model at step $j+1$ as a result of revision ($\mathcal{M}_{i_{j+1}}^{som} = \mathcal{M}_{i_j}^{som} \star C_{i_j}(x_{i_j})$). The knowledge base K (the set of all the strict and defeasible inclusions satisfied in $\mathcal{M}^{som}$) can then be regarded as the knowledge base obtained from K_0 through a sequence of revision steps, i.e., $K = K_0 \star C_{i1}(x_{i1}) \star \ldots \star C_{ir}(x_{ir})$. In fact, from any state of the SOM we can construct a corresponding model, which determines a knowledge base, the set of (strict and defeasible) inclusions satisfied in that model. For future work, it would be interesting to study the properties of this notion of revision and compare with the properties of the notions of iterated belief revision studied in the literature [9, 14, 21, 8].

6 Conclusions

The concept-wise multipreference semantics has recently been introduced for dealing with typicality in description logics [13], based on the idea that reasoning about exceptions in ontologies requires taking into account preferences with respect to different concepts and integrating them into a preferential semantics which allows a standard, KLM style, interpretation of defeasible inclusions.

In this paper, we have explored the relationships between a concept-wise multipreference semantics and self-organising maps. On the one hand, we have seen that self-organising maps can be given a logical semantics in terms of KLM-style preferential interpretations. The model can be used to learn or to validate conditional knowledge from the empirical data used in the category generalization process based on model checking. The learning process in the self-organising map can be regarded as an iterated belief revision process. On the other hand, the plausibility of concept-wise multipreference semantics is supported by the fact that self-organising maps are considered as psychologically and biologically plausible neural network models.

Much work has been devoted, in recent years, to the combination of neural networks and symbolic reasoning. Let us mention Neural Symbolic Computing [11, 10], Logic Tensor Networks [31], and the approaches based on computational logic and logic programming DeepProbLog [26], a probabilistic logic programming language which incorporates deep learning by means of neural predicates, and NeurASP [33], a simple extension of answer set programs that embrace neural networks.

The characterization of self-organising maps in terms of multipreference interpretations, besides providing a logical interpretation to SOMs, which may be of interest from the side of explainable AI, can potentially be exploited, as described above, as a basis for an integrated use of self-organising maps and defeasible knowledge bases.

Acknowledgement: This research is partially supported by INDAM-GNCS Projects 2019 and 2020.

References

1. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In L.P. Kaelbling and A. Saffiotti, editors, *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 364–369, Edinburgh, Scotland, UK, August 2005. Professional Book Center.
2. S. Benferhat, D. Dubois, and H. Prade. Possibilistic logic: From nonmonotonicity to logic programming. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty, European Conference, ECSQARU'93, Granada, Spain, November 8-10, 1993, Proceedings*, pages 17–24, 1993.
3. P. A. Bonatti, M. Faella, I. Petrova, and L. Sauro. A new semantics for overriding in description logics. *Artif. Intell.*, 222:1–48, 2015.
4. Gerhard Brewka. A rank based description language for qualitative preferences. In *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, Valencia, Spain, August 22-27, 2004*, pages 303–307, 2004.
5. K. Britz, J. Heidema, and T. Meyer. Semantic preferential subsumption. In G. Brewka and J. Lang, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 11th International Conference (KR 2008)*, pages 476–484, Sidney, Australia, September 2008. AAAI Press.
6. G. Casini, T. Meyer, I. J. Varzinczak, , and K. Moodley. Nonmonotonic Reasoning in Description Logics: Rational Closure for the ABox. In *26th International Workshop on Description Logics (DL 2013)*, volume 1014 of *CEUR Workshop Proceedings*, pages 600–615, 2013.
7. G. Casini and U. Straccia. Rational Closure for Defeasible Description Logics. In T. Janhunen and I. Niemelä, editors, *Proc. 12th European Conf. on Logics in Artificial Intelligence (JELIA 2010)*, volume 6341 of *LNCS*, pages 77–90, Helsinki, Finland, September 2010. Springer.
8. J. Chandler and R. Booth. Revision by conditionals: From hook to arrow. In *Proc. KR 2020, 17th International Conference on Principles of Knowledge Representation and Reasoning*. AAAI Press, 2020.
9. A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89:1–29, 1997.
10. A. S. d'Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and Son N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6(4):611–632, 2019.
11. A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer, 2009.
12. J. P. Delgrande. A first-order conditional logic for prototypical properties. *Artificial Intelligence*, 33(1):105–130, 1987.
13. L. Giordano and D. Theseider Dupré. An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. *Theory and Practice of Logic programming, TPLP*, 10(5):751–766, 2020. https://doi.org/10.1017/S1471068420000381.
14. L. Giordano, V. Gliozzi, and N. Olivetti. Iterated Belief Revision and Conditional Logic. *Studia Logica*, 70:23–47, 2002.
15. L. Giordano, V. Gliozzi, N. Olivetti, and G. L. Pozzato. Preferential Description Logics. In Nachum Dershowitz and Andrei Voronkov, editors, *Proceedings of LPAR 2007 (14th Conference on Logic for Programming, Artificial Intelligence, and Reasoning)*, volume 4790 of *LNAI*, pages 257–272, Yerevan, Armenia, October 2007. Springer-Verlag.
16. L. Giordano, V. Gliozzi, N. Olivetti, and G. L. Pozzato. Reasoning about typicality in low complexity DLs: the logics $\mathcal{EL}^{\perp}\mathbf{T}_{min}$ and *DL-Lite*$_c\mathbf{T}_{min}$. In *Proc. 22nd Int. Joint Conf. on Artificial Intelligence (IJCAI 2011)*, pages 894–899, Barcelona, July 2011. Morgan Kaufmann.

17. L. Giordano, V. Gliozzi, N. Olivetti, and G. L. Pozzato. Semantic characterization of rational closure: From propositional logic to description logics. *Artificial Intelligence*, 226:1–33, 2015.
18. L. Giordano, V. Gliozzi, N. Olivetti, and G.L. Pozzato. Minimal Model Semantics and Rational Closure in Description Logics . In *26th International Workshop on Description Logics (DL 2013)*, volume 1014, pages 168 – 180, 7 2013.
19. L. Giordano and D. Theseider Dupré. ASP for minimal entailment in a rational extension of SROEL. *TPLP*, 16(5-6):738–754, 2016. DOI: 10.1017/S1471068416000399.
20. V. Gliozzi and K. Plunkett. Grounding bayesian accounts of numerosity and variability effects in a similarity-based framework: the case of self-organising maps. *Journal of Cognitive Psychology*, 31(5–6), 2019.
21. G. Kern-Isberner. A thorough axiomatization of a principle of conditional preservation in belief revision. *Ann. Math. Artif. Intell.*, 40(1-2):127–164, 2004.
22. T. Kohonen, M.R. Schroeder, and T.S. Huang, editors. *Self-Organizing Maps, Third Edition*. Springer Series in Information Sciences. Springer, 2001.
23. S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.
24. D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.
25. D. Lewis. *Counterfactuals*. Basil Blackwell Ltd, 1973.
26. R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 3753–3763, 2018.
27. R. Miikkulainen, J. Bednar, Y. Choe, and J. Sirosh. *Computational maps in the visual cortex*. Springer, 2002.
28. D. Nute. Topics in conditional logic. *Reidel, Dordrecht*, 1980.
29. J. Pearl. System Z: A Natural Ordering of Defaults with Tractable Applications to Nonmonotonic Reasoning. In R. Parikh, editor, *TARK (3rd Conference on Theoretical Aspects of Reasoning about Knowledge)*, pages 121–135, Pacific Grove, CA, USA, 1990. Morgan Kaufmann.
30. M. Pensel and A. Turhan. Reasoning in the defeasible description logic $EL_\perp$ - computing standard inferences under rational and relevant semantics. *Int. J. Approx. Reasoning*, 103:28–70, 2018.
31. Luciano Serafini and Artur S. d'Avila Garcez. Learning and reasoning with logic tensor networks. In *AI*IA 2016: Advances in Artificial Intelligence - XVth Int. Conf. of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 - December 1, 2016, Proceedings*, volume 10037 of *LNCS*, pages 334–348. Springer.
32. J. B. Tenenbaum and T. L. Griffiths. Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, 24:629–641, 2001.
33. Z. Yang, A. Ishay, and J. Lee. Neurasp: Embracing neural networks into answer set programming. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1755–1762. ijcai.org, 2020.