Towards a Conditional Interpretation of Self Organizing Maps ⋆

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

[1] DISIT - Università del Piemonte Orientale, Alessandria, Italy
[2] Center for Logic, Language and Cognition & Dipartimento di Informatica, Università di Torino, Italy,

Abstract. In this paper we aim at establishing a link between the preferential semantics for conditionals and self-organising maps (SOMs). We show that a concept-wise multipreference semantics, recently proposed for defeasible description logics, which takes into account preferences with respect to different concepts, can be used to to provide a logical interpretation of SOMs.

1 Introduction

Preferential approaches [15, 16] to common sense reasoning, having their roots in conditional logics [17, 19], have been 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 [10, 3] and closure constructions [5, 4, 12, 20].

In this paper we study the relationships between preferential semantics for conditionals and self-organising maps (SOMs)[14], psychologically and biologically plausible neural network models that can learn after limited exposure to positive category examples, without any need of contrastive information. Self-organising maps have been proposed as possible candidates to explain the psychological mechanisms underlying category generalisation.

We show that a "concept-wise" multipreference semantics [8], recently proposed for a lightweight description logic of the $\mathcal{EL}^{\perp}$ family, can be used to provide a logical semantics of SOMs. The result of the process of category generalization in self-organising maps can be regarded as a multipreference model in which different preference relations are associated to different concepts (the learned categories). The combination of these preferences into a global preference, following the approach in [8], defines a standard KLM preferential interpretation. Such an interpretation can be used to learn or validate conditional knowledge from the empirical data used in the category generalization process. The evaluation of conditionals can be done by model checking, using the information recorded in the SOM. We believe that the proposed semantic interpretation of SOMs can be relevant in the context of explainable AI.

These results have been first presented at CILC 2020 [9].

2 A concept-wise multi-preference semantics

In this section we shortly describe an extension of $\mathcal{EL}^\bot$ with typicality inclusions, defined along the lines of the extension of description logics with typicality [10, 11], and its multi-preference semantics [8].

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 $a, b \in N_I$.

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 [15, 16]. 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 distinguished $\mathcal{EL}^\bot$ 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 [16], i.e., it is a modular and well-founded strict partial order (an 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 1 (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, as in an* $\mathcal{EL}^\bot$ *interpretation 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 preference relation $<_{C_i}$ allows the set of prototypical C_i-elements to be defined as the C_i-elements which are minimal with respect to $<_{C_i}$, i.e., $min_{<_{C_i}}(C_i^I)$. 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 [8], which have been inspired by Brewka's framework of basic preference descriptions [2]. While we refer to [8] for the construction of the preference relations $<_{C_i}$'s from a ranked knowledge base K, in the following we shortly 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, starting from a ranked KB containing inclusions $\mathbf{T}(Stud) \sqsubseteq Young$, $\mathbf{T}(Emp) \sqsubseteq Has\_Boss$, $\mathbf{T}(Emp) \sqsubseteq NonYoung$, and $Young \sqcap NonYoung \sqsubseteq \bot$. To answer the questions above both preference relations $<_{Emp}$ and $<_{Stud}$ are relevant, and they might be conflicting as, for instance, Tom is more typical than Bob as a student ($tom <_{Stud} bob$), but more exceptional as an employee ($bob <_{Emp} tom$). By *combining* the preference relations $<_{C_i}$ into a single *global preference* relation $<$ we can exploit the global preference $<$ for interpreting the typicality operator, which may be applied to arbitrary concepts, and verify, for instance, whether $\mathbf{T}(Stud \sqcap Emp) \sqsubseteq Has\_Boss$.

A natural definition of the notion of global preference $<$ exploits Pareto combination of the relations $<_{C_1}, \ldots, <_{C_k}$, as follows:

$$x < y \text{ iff } \quad (i)\ x <_{C_i} y, \text{ for some } C_i \in \mathcal{C}, \text{ and}$$
$$(ii)\ \text{ for all } C_j \in \mathcal{C},\ x \leq_{C_j} y$$

where $\leq_{C_i}$ is the non-strict preference relation associated with $<_{C_i}$ ($\leq_{C_i}$ is a total preorder). A slightly more sophisticated notion of preference combination, which exploits a modified Pareto condition taking into account the specificity relation among concepts (such as, for instance, the fact that concept $PhdStudent$ is more specific than concept $Student$), has been considered for ranked knowledge bases [8].

The addition of the global preference relation allows for defining a notion of *conceptwise multipreference interpretation* $\mathcal{M} = \langle \Delta, <_{C_1}, \ldots, <_{C_k}, <, \cdot^I \rangle$, where typicality concept $\mathbf{T}(C)$ is interpreted as the set of the $<$-minimal C elements, i.e., $(\mathbf{T}(C))^I = min_<(C^I)$, where $Min_<(S) = \{u : u \in S \text{ and } \nexists z \in S \text{ s.t. } z < u\}$.

The notions of cw^m-model of a ranked $\mathcal{EL}^{\bot}$ knowledge base K, and of cw^m-entailment can be defined in the natural way. In particular, cw^m-entailment has been proved to satisfy the KLM postulates of a preferential consequence relation [8].

3 Self-organising maps

Self-organising maps (SOMs, introduced by Kohonen [14]) are particularly plausible neural network models that learn in a human-like manner. 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 [13]. Roughly speaking, in [13] 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 [14]. Each map unit u is associated with a weight vector w_u of the same dimensionality 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. The learning process is incremental: after the presentation of each input, the map's representation of the input (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 [13] 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 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) [13] :

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

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.

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 [13] that the Numerosity and Variability effects of category generalization, described by Griffiths and Tenenbaum [22], 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.

4 Relating self-organising Maps and multi-preference models

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, 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 the space of input stimuli to be 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 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 concept-wise multipreference interpretation associated with the SOM. We restrict the DL language to the fragment of $\mathcal{EL}^{\bot}$ (plus typicality) not admitting roles.

Definition 2 (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:*

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

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 in the fragment of* $\mathcal{EL}^{\bot}$ *without roles according to Definition 1.*

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. As, for $y = BMU_x$, $rd(y, C_i)$ is 0, $BMU_{C_i} \subseteq min_{<_{C_i}}(C_i^I)$.

4.1 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), by exploiting a notion of maximal relative distance of BMU_{C_i} from C_j, defined as $rd(BMC_{C_i}, C_j) = max_{x \in C_i}\{rd(BMU_x, C_j)\}$.

While we refer to [9] for details, let us observe that checking the satisfiability of strict or defeasible inclusions on the SOM may be non trivial, depending on the number of input stimuli that have been considered in the learning phase, although 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 [13].

Note also that the multipreference interpretation $\mathcal{M}^s$ introduced in Definition 2 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 $<_{C_i}$. Although, we are not able to define the most typical $C_i \sqcap C_j$-elements just using single preferences. Starting from $\mathcal{M}^s$, we can 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 $\mathcal{M}^{som}$ can be constructed from $\mathcal{M}^s$ according to the definition of the global preference in Section 2.

We have focused on the multipreference interpretation of a self-organising map after the learning phase. However, the state of the SOM during the learning phase can as well be represented as a multipreference model (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). We can than regard the category generalization process as a model building process and, in a way, as a belief revision process.

5 Conclusions

We have explored the relationships between a concept-wise multipreference semantics and self-organising maps, showing that conditional logics can be used to provide a logical explanation to self-organising maps. In particular, self-organising maps can be given a logical semantics in terms of KLM-style preferential interpretations. In particular, 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.

Much work has been devoted, in recent years, to the combination of neural networks and symbolic reasoning. Let us mention Neural Symbolic Computing [7, 6], Logic Tensor Networks [21], and the approaches based on computational logic and logic programming DeepProbLog [18], a probabilistic logic programming language which incorporates deep learning by means of neural predicates, and NeurASP [23], a simple extension of answer set programs that embrace neural networks.

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.
2. G. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer, 2009.
8. 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.
9. L. Giordano, V. Gliozzi, and D. Theseider Dupré. On a plausible concept-wise multipreference semantics and its relations with self-organising maps. *CoRR*, abs/2008.13278, 2020. To appear in CILC (Italian Conference on Computational Logic), 13-15 October 2020, Rende.
10. 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.
11. 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.
12. 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.
13. 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.
14. T. Kohonen, M.R. Schroeder, and T.S. Huang, editors. *Self-Organizing Maps, Third Edition*. Springer Series in Information Sciences. Springer, 2001.
15. S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.
16. D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.
17. D. Lewis. *Counterfactuals*. Basil Blackwell Ltd, 1973.

18. 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.
19. D. Nute. Topics in conditional logic. *Reidel, Dordrecht*, 1980.
20. 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.
21. L. Serafini and A. 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.
22. J. B. Tenenbaum and T. L. Griffiths. Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, 24:629–641, 2001.
23. 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.