Learning fuzzy concept inclusions from OWL real-valued data

F.A. Cardillo, *ILC-CNR, Italy*
francoalberto.cardillo@ilc.cnr.it

U.Straccia, *ISTI-CNR, Italy*
umberto.straccia@isti.cnr.it

OWL Data &
Real-valued attributes

Goal:

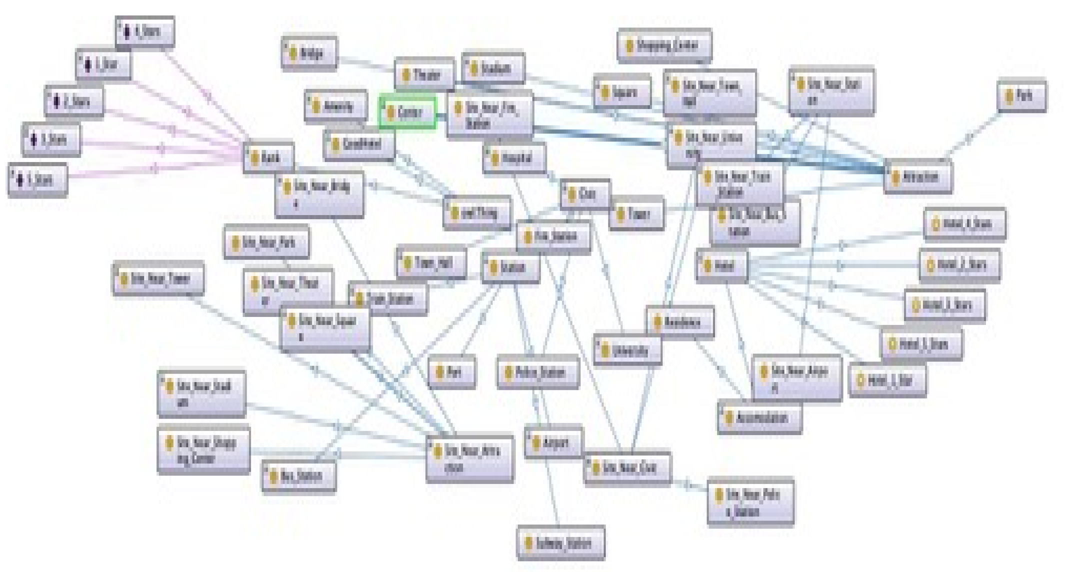
Learn OWL EL descriptions of a target class T

Reasoning with:
OWL & Fuzzy OWL
Fuzzy Classes via Data Clustering

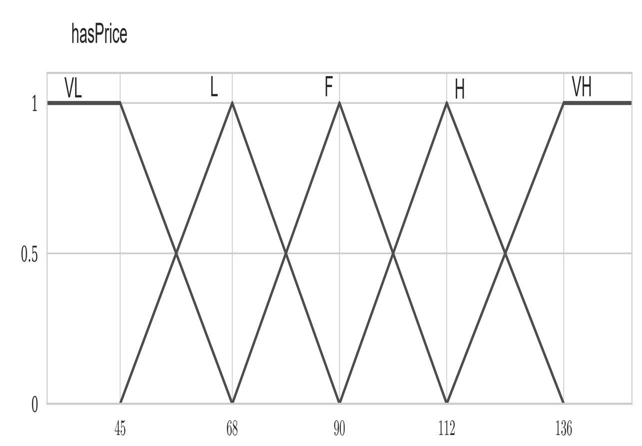
$\Rightarrow$ Learnt descriptions of T is easy to interpret

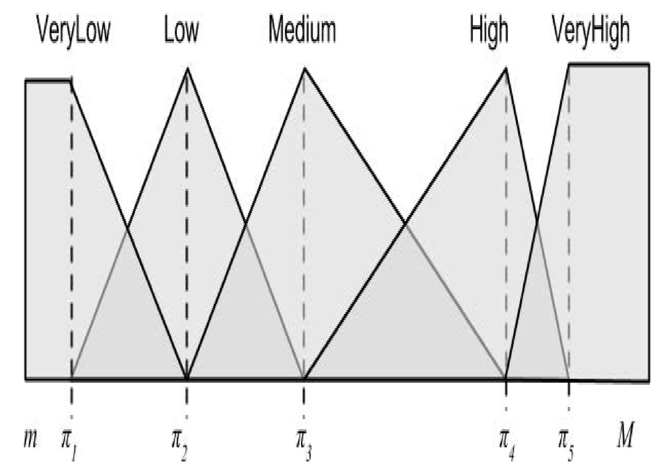
OWL Learning Algorithms:
Fuzzy Foil
Fuzzy Real AdaBoost

Given a crisp OWL 2 ontology and a target concept T, we address the problem of learning sufficient conditions for an individual to be an instance of T.



Real-valued properties, such as, e.g., the price of a room, are mapped onto fuzzy sets either via partitioning their range into intervals with equal width or via clustering, using the c-means algorithm, where the membership functions are triangular and centred on the centroids of the clusters.





We adapted Real AdaBoost to fuzzy OWL ontologies $\mathcal{K}$:

- incremental learning of single axioms;
- the weak learner returns set of axioms, covering as many positive examples as possible (parametric);
- such sets are linearly combined by our modified Real Adaboost.

Fuzzy OWL-Boost

Input: KB $\mathcal{K}$, training set $\mathcal{E}$, target concept name T, number of iterations n

$h \leftarrow \emptyset$, $\mathsf{I} \leftarrow \mathsf{I}_{\mathcal{K}}$;
$\mathbf{w}_1 \leftarrow \mathbf{u}$;
for $i = 1$ to n **do**
  $h_i \leftarrow \textsc{FuzzyWeakLearner}(\mathcal{K}, T, \mathcal{E}, \mathbf{w}_i)$;
  if $\epsilon(\mathbf{w}_i) \geq 0.5$ **then break**;
  $h_i^* \leftarrow \max_{a \in \mathsf{I}} |h_i(a)|$;
  $\mu_i \leftarrow \frac{1}{h_i^*} \Sigma_{a \in \mathsf{I}}\, w_{i,a} \cdot l(a) \cdot h_i(a)$;
  $\alpha_i \leftarrow \frac{1}{2h_i^*} \cdot \ln \frac{1+\mu_i}{1-\mu_i}$;
  for all $a \in \mathsf{I}$ **do**
    $w_{i+1,a} \leftarrow w_{i+1,a} \cdot \left(\frac{1-(\mu_i \cdot l(a) \cdot h_i(a))/h_i^*}{1-\mu_i^2}\right)$;
  $h \leftarrow h \cup \{C_{ij} \sqsubseteq WL_i \mid C_{ij} \sqsubseteq T \in h_i, WL_i \text{ new}\}$;
$\phi_T \leftarrow \alpha_1 \cdot WL_1 + \ldots + \alpha_n \cdot WL_n \sqsubseteq T$;
$h \leftarrow h \cup \{\phi_T\}$;
return h;

- Weak learner h_i, *i.e.* set of a axioms $C_{ij} \sqsubseteq T$.
- The error of the weak learner needs to be < 0.5. If not, break the loop.
- μ_i is the *normalised* margin of weak learner h_i w.r.t. I.
- α_i will weight the prediction of the classifier h_i in the final ensemble.
- Update the weight distribution, increasing the weights of wrongly classified examples and decreasing the weights of those correctly classified.
- Update hypothesis. WL_i are new atomic symbols.
- Build now the final classifier ensemble as a weighted (linear) combination of the weak learners.
- Include h since it keeps the atomic symbols:$\{WL_k\}$

Results. Experiments: 15 ontologies (some from UCI Machine Learning Repository): Fuzzy OWL-Boost better than Fuzzy Foil-$\mathcal{DL}$

Dataset, Target	Algorithm	Clustering	theta	fs	MSE	Fuzzy F1	F1	Best (FuzzyF1 * F1)	% (Best)
Iris (setosa)	FOIL-DL	u	0.34	3	0.018	0.893	1.000	0.893	
	Fuzzy OWL-Boost	u	0.34	3	0.000	1.000	1.000	1.000	11.93%
Iris (versicolor)	FOIL-DL	c	0.94	7	0.090	0.710	0.888	0.630	
	Fuzzy OWL-Boost	c	0.64	3	0.021	0.929	0.860	0.799	26.68%
Iris (virginica)	FOIL-DL	u	0.94	5	0.048	0.851	0.891	0.758	
	Fuzzy OWL-Boost	u	0.64	5	0.036	0.922	0.891	0.821	8.30%
Wine (1)	FOIL-DL	u	0.94	5	0.076	0.742	0.915	0.679	
	Fuzzy OWL-Boost	u	0.94	3	0.049	0.897	0.907	0.813	19.79%
Wine (2)	FOIL-DL	c	0.94	7	0.074	0.837	0.885	0.741	
	Fuzzy OWL-Boost	c	0.94	7	0.083	0.844	0.910	0.769	3.75%
Wine (3)	FOIL-DL	c	0.64	7	0.033	0.868	0.885	0.768	
	Fuzzy OWL-Boost	u	0.94	7	0.017	0.955	0.932	0.890	15.86%
Wine Quality	FOIL-DL	c	0.94	7	0.003	0.034	0.053	0.002	
	Fuzzy OWL-Boost	c	0.04	5	0.011	0.047	0.065	0.003	68.49%
FamilyTree	FOIL-DL	u/c	0.94	5, 3	0.019	0.929	0.929	0.863	
	Fuzzy OWL-Boost	u/c	0.94	3	0.018	0.924	0.924	0.853	-1.20%
Hotel (*)	FOIL-DL	c	0.94	7	0.007	0.942	1.000	0.942	
	Fuzzy OWL-Boost	c	0.34	7	0.005	0.968	1.000	0.968	2.76%
Moral	FOIL-DL	-	0.34	3	0.000	1.000	1.000	1.000	
	Fuzzy OWL-Boost		0.34	3	0.000	1.000	1.000	1.000	0.00%
SemanticBible (NTN)	FOIL-DL	u/c	0.34	3	0.046	0.527	0.548	0.289	
	Fuzzy OWL-Boost	u/c	0.94	3	0.040	0.541	0.541	0.293	1.46%
UBA (*)	FOIL-DL	c	0.34	3	0.001	0.972	0.977	0.950	
	Fuzzy OWL-Boost	c	1.00	3	0.000	1.000	1.000	1.000	5.30%
Pair50 (*)	FOIL-DL	-	0.34	3	0.000	1.000	1.000	1.000	
	Fuzzy OWL-Boost	-	0.34	3	0.000	1.000	1.000	1.000	0.00%
Straight (*)	FOIL-DL	-	0.34	3	0.007	0.400	0.571	0.228	
	Fuzzy OWL-Boost	-	0.34	3	0.021	0.514	0.500	0.257	12.52%
WineOnto (*)	FOIL-DL	-	0.34	3	0.000	1.000	1.000	1.000	
	Fuzzy OWL-Boost	-	0.34	3	0.000	1.000	1.000	1.000	0.00%
Lymphography	FOIL-DL	-	0.34	3	0.159	0.845	0.855	0.722	
	Fuzzy OWL-Boost	-	0.64	3	0.160	0.861	0.870	0.749	3.71%
Mammographic	FOIL-DL	u	0.34	3	0.222	0.593	0.710	0.421	
	Fuzzy OWL-Boost	c	0.64	7	0.199	0.720	0.774	0.557	32.26%
Pyrimidine (*)	FOIL-DL	u	1.00	7	0.050	0.889	0.947	0.842	
	Fuzzy OWL-Boost	c	0.34	7	0.027	0.952	0.930	0.885	5.16%
Suramin (*)	FOIL-DL	c	0.34	3	0.002	0.291	0.583	0.170	
	Fuzzy OWL-Boost	u/c	0.34	3	0.003	0.583	0.583	0.340	100.34%

Learnt descriptions can be easily turned into natural language:

- Hotel (based on Tripadvisor data about hotels in Pisa):

  "An expensive B&B offering a cradle and Wi-Fi is a good hotel".

  `Bed_and_Breakfast and (hasAmenity some Cradle) and (hasAmenity some WI−FI) and ( hasPrice some hasPrice_high) SubClassOf GoodHotel`

- Mammography (based on BI-RADS attributes and the patient's age):

  "An old woman having a high density mass, whose margin is obscured and that has an irregular shape has a malignant mammographic mass".

  `(hasAge some hasAge_high) and (hasDensity some hasDensityHigh) and (hasMargin some obscured) and (hasShape some irregular) SubClassOf MalignantMammographicMass`

References. F.A. Cardillo, U. Straccia. "Fuzzy OWL-Boost: Learning fuzzy concept inclusions via real-valued boosting. To appear in "*Fuzzy Sets and Systems*", 2021.

Partially supported by the projects TAILOR (GA No 952215) and DeepHealth (GA No. 825111).