CARAF: Complex Aggregates within RAndom Forests

Clément Charnay, Nicolas Lachiche, Agnès Braud

ICube, Université de Strasbourg, CNRS
300 Bd Sébastien Brant - CS 10413
F-67412 Illkirch Cedex

{charnay,nicolas.lachiche,agnes.braud}@unistra.fr

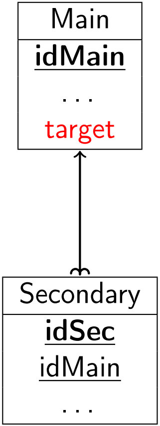
ILP 2015
Kyoto, 2015/08/22

Outline

Relational Data

Relational Data

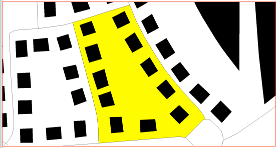
- Data represented across several tables: different kinds of objects.
- In this work, 2 tables:
  - main: objects we want to predict on,
  - secondary: objects in 1-to-many relationship with main table, composition for instance.
- Could be a star schema: one main table with several secondary tables directly related to the main.
- Task: build features of main objects using properties of secondary objects



Urban Blocks

Urban Blocks Dataset

- Urban blocks composed of several buildings.
- Learning task : predict of which kind is the urban block according to geometric properties.
- Data available from former project: 591 urban blocks from 4 areas of Strasbourg, composed of 7692 buildings.
- 6 class-problem.



Urban Blocks Example - Schema

block_id	density	convexity	elongation	area	class
1	0.151	0.986	0.221	22925	h_indiv
2	0.192	0.832	0.155	15363	h_coll
3	0.204	0.718	0.450	17329	h_mixed
...	...	...	...	...	...

Main table - blocks

1

0..N

building_id	convexity	elongation	area	block_id
1_1	1.000	0.538	165	1
1_2	0.798	0.736	323	1
1_3	1.000	0.668	84	1
...	...	...	...	...
2_1	0.947	0.925	202	2
2_2	1.000	0.676	147	2
...	...	...	...	...

Secondary table - buildings

State of the Art

Possible Approaches

- Tilde[a]: logical decision tree induction, introduction of secondary objects through existential quantifier.
- RELAGGS[b]: propositionalization through simple aggregation.

[a] Hendrik Blockeel and Luc De Raedt. "Top-Down Induction of First-Order Logical Decision Trees". In: *Artif. Intell.* 101.1-2 (1998), pp. 285–297.

[b] M.-A. Krogel and S. Wrobel. "Facets of Aggregation Approaches to Propositionalization". In: *Work-in-Progress Track at the Thirteenth International Conference on Inductive Logic Programming (ILP)*. 2003.

Our aim

- Introduce relevant secondary objects (like Tilde).
- Use aggregation to go further than the existential quantifier.

$\Rightarrow$ **Complex Aggregation**

Outline

Complex Aggregates - Introduction

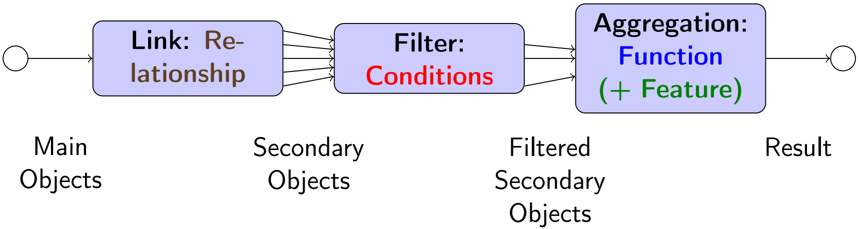
What is a Complex Aggregate

- Constructed feature of the objects of the main table.
- Aggregates the values of a feature of secondary objects that meet a certain condition.

Composition of Complex Aggregates

- Selection of secondary objects:
  - Link: Relationship between tables.
  - Filter: Conditions on secondary objects.
- Aggregation process:
  - Attribute to aggregate (not always).
  - Aggregation function.

Examples of Complex Aggregates



Examples

- Number of buildings in the block.
- Maximum area of buildings with elongation $\geq$ 0.5.
- Average elongation of buildings with convexity $<$ 0.9 and area $\geq$ 150.

Example - Notation

avg(elongation, buildings, convexity $< 0.9 \wedge$ area ≥ 150)

Searching the Feature Space

Explosion of Search Space

- Problem: number of complex aggregates for a given problem is combinatorial, impossible to consider them all!
- Especially, the aggregation condition is a conjunction of several basic conditions.

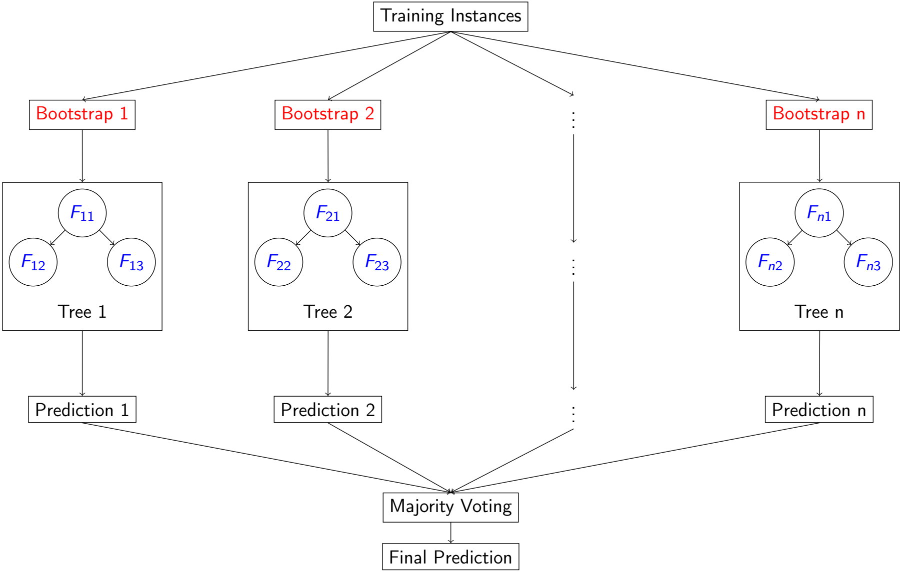
ILP 2014

- RRHCCA[1]: Random Restart Hill-Climbing of Complex Aggregates.
- In a **single decision tree**, find splits on complex aggregates.
- Given the aggregation process, find the best conjunction of conditions through random restart hill-climbing.

[1] C. Charnay, N. Lachiche, and A. Braud. "Construction of Complex Aggregates with Random Restart Hill-Climbing". In: *24th International Conference on Inductive Logic Programming (ILP'14)*. 2014.

Outline

Random Forests



Random Forests for Complex Aggregates

Motivation

- Large Feature Space. ($\approx F \cdot A \cdot N^A$)
- Complex aggregates are specific, overfitting with a single decision tree.
- Relax the optimization method to search through the feature space.

Existing Methods

- Tilde extended to both complex aggregates and Random Forests, FORF[a].
- However, memory problems when language bias allows big conjunction for selection condition.
- Feature sampling is uniform $\rightarrow$ may not create enough diversity.

[a] Anneleen Van Assche et al. "First order random forests: Learning relational classifiers with complex aggregates". In: *Machine Learning* 64.1-3 (2006), pp. 149–182.

Random Forests in CARAF

Complex Aggregate Feature Sampling

- Bootstrapping and recombination are classic.[2]
- Structural feature sampling: keep square root of aggregation processes and half of the attributes for conditions. (sampled feature space size $\approx$ square root of the original feature space size)

Func \ Attr	Area	Elong	Conv
Average	x		
Min			
Max			
Std Dev		x	
Sum	x		

Cond	Area	Elong	Conv
	x	x	

[2] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

Hill-Climbing of Aggregation Conditions

Hill-Climbing Strategies

- RRHCCA (ILP 2014): given the aggregation process, find the best conjunction by testing a neighborhood of refinements at each step.
- Random: given the aggregation process, try one random neighbor at each step.
- Global: try one random neighbor condition at each step, on every aggregation process at hand.

Refinements

From original condition $\text{area} \geq 150$, we can refine to:

- Empty condition.
- $\text{area} \geq 150 \wedge \text{elongation} < 0.6$
- $\text{area} \geq 120$
- $\text{area} \geq 180$

Outline

Out-of-bag Accuracy Results

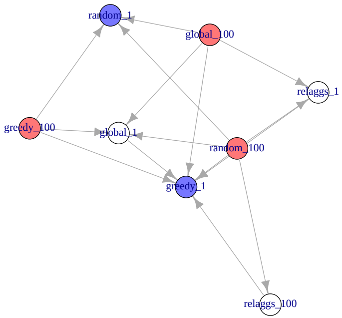
Out-of-bag Accuracy

- For each training instance, use sub-forest that did not see the instance at training to classify.
- Used to compare Random Forests.
- 33 trees in each forest.

Dataset	RELAGGS	FORF	RRHCCA	Random	Global
Auslan	94.19%	ERR	**96.53%**	**95.91%**	94.66%
Diterpenes	89.09%	90.49%	<u>**92.95%**</u>	85.06%	<u>**93.35%**</u>
Jp-Vowels	93.78%	94.86%	95.41%	**97.30%**	**97.03%**
Musk1	80.43%	78.26%	<u>89.13%</u>	84.78%	80.43%
Musk2	76.47%	75.49%	81.37%	85.29%	82.35%
Opt-digits	22.37%	76.57%	<u>**95.94%**</u>	<u>**94.60%**</u>	<u>**92.77%**</u>
Urban	83.42%	75.81%	<u>84.94%</u>	<u>83.76%</u>	<u>84.60%</u>
			7 - 6	6 - 5	6.5 - 6

10-fold Cross Validation Results

Dataset	Muta	Urban
RELAGGS-1	89.40%	74.86%
RELAGGS-100	90.26%	84.55%
RRHCCA-1	84.86%	74.69%
RRHCCA-100	91.33%	87.48%
Random-1	87.67%	75.55%
Random-100	92.22%	87.28%
Global-1	87.82%	74.60%
Global-100	91.96%	87.68%



Conclusion and Future Work

Conclusion

- Random Forests improve over Decision Trees with Complex Aggregates.
- Our Hill-Climbing algorithms perform better than RELAGGS and FORF.
- Faster hill-climbing algorithms do not yield loss of accuracy.

Future Work

- Do Feature Selection with Random Forests: find most relevant families of aggregates.
- Handle Nested Relationships, especially complex aggregates as aggregated feature.