

Rule-Based Classification Approach: Closed Itemsets vs Random Forests

Tatiana Makhalova

October 20 2017,
National Research University Higher School of Economics,
Moscow

Random Forests or Deep Networks?

Random Forest **or and** Deep Network

- ▶ Unsupervised Pretrained Networks with supervised-trained model on the top (Bengio, 2009).
- ▶ Distributed algorithms for RF and neural-tree RF (Yildiz and Alpaydin, 2013).
- ▶ Networks in the pretraining stage of RF learning (Kontschieder et al. 2015, Ioannou et al. 2016).


What Supervised Model to Choose?

- ▶ Comparison of 179 classifiers from 17 families¹ on whole UCI data: “the classifiers **most likely to be the best** are the random forest”.

Fernandez-Delgado, Manuel, et al. “Do we need hundreds of classifiers to solve real world classification problems.” J. Mach. Learn. Res 15.1 (2014): 3133-3181.

- ▶ Revision of the results: “random forests **do not have significantly higher percent accuracy** than support vector machines and neural networks, calling into question the conclusion that random forests are the best classifiers”.

Wainberg, Michael, Babak Alipanahi, and Brendan J. Frey. “Are random forests truly the best classifiers?.” The Journal of Machine Learning Research 17.1 (2016): 3837-3841.

¹discriminant analysis, Bayesian, NN, SVM, decision trees, rule-based classifiers, boosting, bagging, stacking, RF, generalized linear models, nearest neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and others 

Domain-based Analysis

- ▶ **Chemistry** “...RF typically yields comparable or possibly better predictive performance than the linear modeling approaches” .

Marchese Robinson, Richard L., et al. "Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets." Journal of Chemical Information and Modeling 57.8 (2017): 1773-1792.

- ▶ **Engineering:** “...both SVM and RFR are excellent choices for electrical load forecast (parameter and data dependent models)” .

Huo, Juan, Tingting Shi, and Jing Chang. "Comparison of Random Forest and SVM for electrical short-term load forecast with different data sources." Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on. IEEE, 2016.

- ▶ **Geology, landslide susceptibility:** “... RF, BRT, CART, and GLM models produced reasonable accuracy in landslide susceptibility mapping” .

Youssef, Ahmed Mohamed, et al. "Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia." Landslides 13.5 (2016): 839-856. ▶

Tree Ensemble. Different Kinds



Bagging: build classifiers on randomly selected subsets of objects.

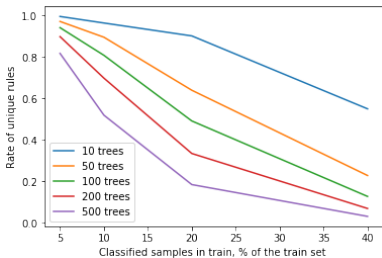
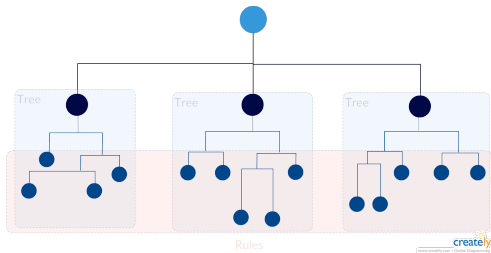
Random Forest: build classifiers on the whole dataset described by random subset of attributes.

Random Forest with bootstrap: build classifiers on randomly selected subsets of objects described by random subset of attributes.

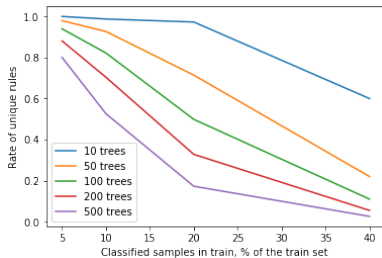
A lot of parameters have to be tuned!

How Many Rules are Unique?

Standard RF-based classification

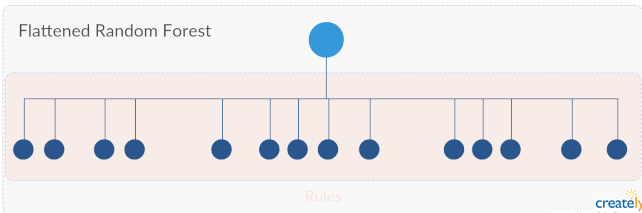
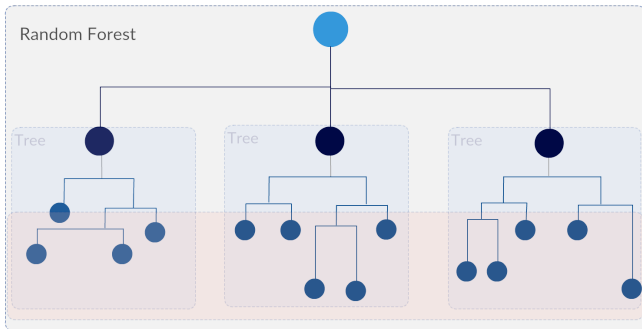


synthetic, 5 informative, 15
random attributes



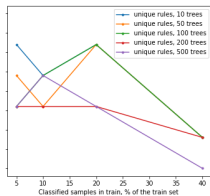
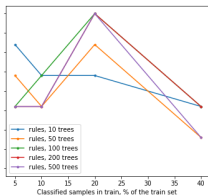
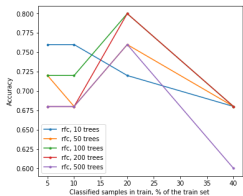
synthetic, 15 informative, 5
random attributes

Random Forests (RF) vs Flattened Random Forests (FRF)

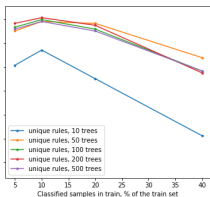
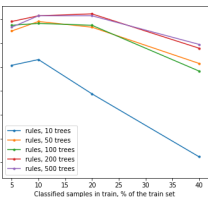
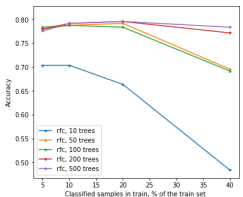


RF, FRF or Unique Rules in FRF?

Comparison of Model Accuracy



100 samples (70 training / 30 test set), 15 informative attributes, 5 random



100 samples (70 training / 30 test set), 5 informative attributes, 15 random

RF, FRF or Unique Rules in FRF?

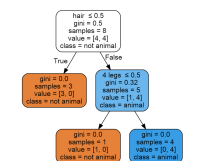
On average,

- ▶ RF and FRF have similar accuracy.
- ▶ RF performs better than FRFs when the number of uninformative attributes is much less than the number of informative ones.
- ▶ Unique-rule-based classifier performs worse than RF and FRF.

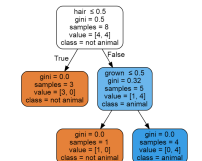
The repeating rules learned by subsamples reveal some structural information from the data rather than relationships between attributes and targets.

Random Forest: An Example

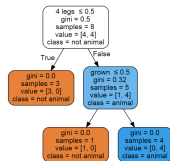
	4 legs	wool	yellow-brown	grown	black-white	target: animal
Sphinx cat	X	X		X	X	+
Dog	X	X		X	X	+
Cat	X	X	X	X		+
Leopard	X	X	X	X		+
Fur coat		X	X			
Chair	X		X			
Sunflower				X		
Ballon				X	X	



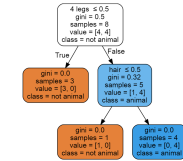
not 'hair' → not animal
 'hair' and not '4 legs' → not animal
 'hair' and '4 legs' → animal



not 'hair' → not animal
 'hair' and 'grown' → animal
 'hair' and not 'grown' → not animal



not '4 legs' → not animal
 '4 legs' and not 'grown' → not animal
 '4 legs' and 'grown' → animal



not '4 legs' → not animal
 '4 legs' and 'hair' → animal
 '4 legs' and not 'hair' → not animal

Simplifying Rule Structure

Can we compress several rules given by the random forest?

“Animal” class rules:

- ▶ 4 legs, grown \rightarrow animal
- ▶ 4 legs, hair \rightarrow animal
- ▶ hair, grown \rightarrow animal
- ▶ 4 legs, hair, grown \rightarrow animal

as well as

- ▶ hair, yellow-brown \rightarrow animal; 4 legs, black-white \rightarrow animal
- ▶ grown, yellow-brown \rightarrow animal; hair, black-white \rightarrow animal
- ▶ ...

We need to define similarity on rules.

Formal Concept Analysis. Basic Notions

A formal context is a triple (G, M, I) , where G is a set objects, M is a set attributes, $I \subseteq G \times M$ is a relation called *incidence relation*. The derivation operators $(\cdot)'$:

$$A' = \{m \in M \mid \forall g \in A : glm\}$$

$$B' = \{g \in G \mid \forall m \in B : glm\}$$

A (formal) concept is a pair (A, B) , where $A \subseteq G$, $B \subseteq M$ and $A' = B$, $B' = A$.

Note that B part of a formal concept is a **closed itemset**, well-known in data mining.

Formal concepts ordered by **generality relation**

$(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2$ make a lattice, called **concept lattice**.

Formal Concepts. Example

	4 legs	wool	yellow-brown	grown	black-white	target: animal
Sphinx cat	X	X		X	X	+
Dog	X	X		X	X	+
Cat	X	X	X	X		+
Leopard	X	X	X	X		+
Fur coat		X	X			
Chair	X		X			
Sunflower				X		
Ball				X	X	

Examples of concepts:

- ▶ $(\{\text{Sphinx cat, Dog, Cat, Leopard}\}, \{4 \text{ legs, wool, grown}\})$
- ▶ $(\{\text{Cat, Leopard}\}, \{4 \text{ legs, wool, yellow-brown, grown}\})$

The Structure on Attribute Space

Lattice vs Tree

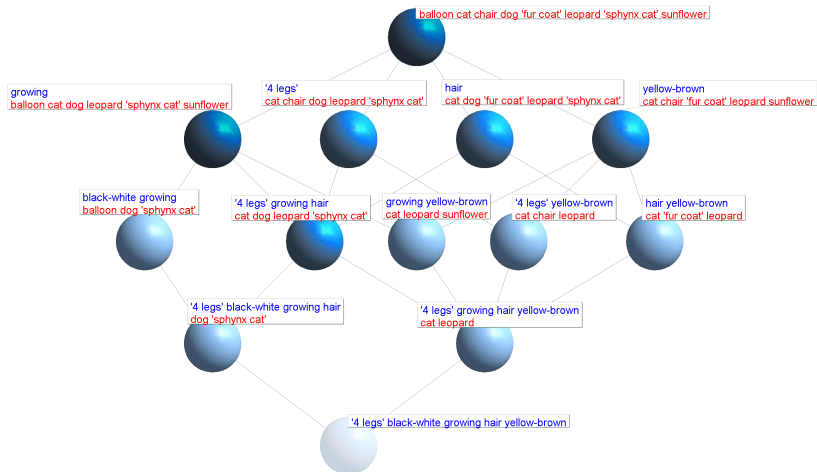


Figure: A lattice of the context "Animals"

The Structure on Attribute Space

Lattice vs Tree: 4 trees in 1 concept

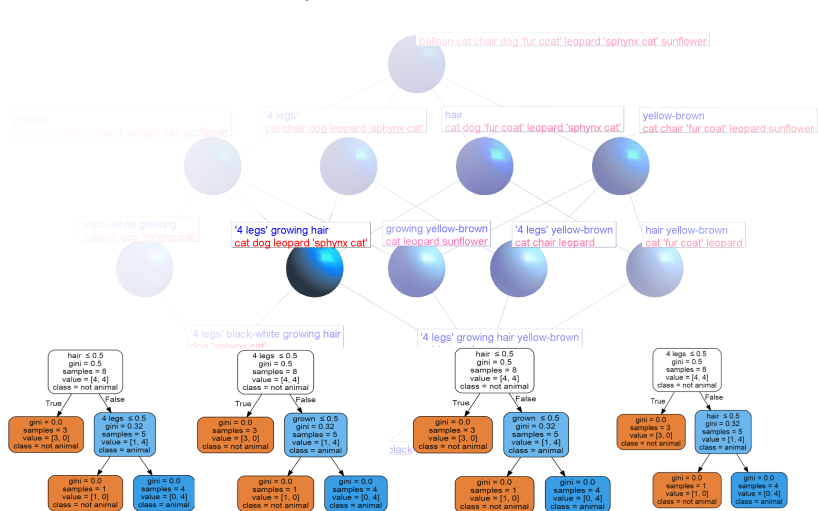


Figure: A lattice of the context "Animals"

Association Rules

$A \rightarrow_{s,c} B$, where $A, B \subseteq M$ holds in context (G, M, I) if

$$\text{support } s = \frac{(A \cup B)'}{|A'|}$$

$$\text{confidence } c = \frac{(A \cup B)'}{|G'|}.$$

All association rules of a dataset are represented concisely by the lattice diagram (Luxenburger basis), [Lakhal et al., 2000]

Implications (Exact Associative Rules)

Example

	4 legs	wool	yellow-brown	grown	black-white	target: animal?
Sphinx cat	X	X		X	X	+
Dog	X	X		X	X	+
Cat	X	X	X	X		+
Leopard	X	X	X	X		+
Fur coat		X	X			
Chair	X		X			
Sunflower				X		
Ball				X	X	

Implications:

black-white \rightarrow grown

4 legs, wool \rightarrow grown

Associative rules:

4 legs \rightarrow wool, confidence = $4/5$

yellow-brown \rightarrow grown, confidence = $1/2$

Concept-based Classifiers

Training stage:

compute formal concepts $\mathcal{S} = \{C = (A, B)\}$ on training set

$G = G_+ \cup G_-$;

assign classes $class(C) = \mathit{argmax}_{s \in \{+, -\}} \left(\frac{|A_s|}{|A|} \right)$ (dominating class).

Test stage:

For each object g and its attribute set g' :

$n_s = |\{B \subseteq g' \mid (A, B) \in \mathcal{S}, class((A, B)) = s\}|$;

$class(g) = \mathit{argmax}_{s \in \{+, -\}} \left(\frac{n_s}{|\{C \in \mathcal{S} \mid class(C) = s\}|} \right)$.

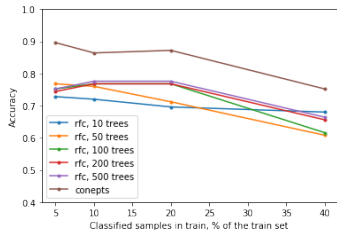
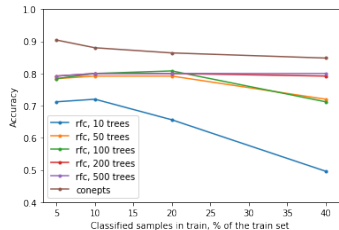
Concept-based Classifiers vs Random Forests

- ▶ **Datasets:** 100 objects, 20 attributes, two classes²;
- ▶ **4 batches of experiments:** the number of informative / random attributes: 20/0, 15/5, 10/10, 5/15.
- ▶ **Each batch:** 100 random splittings, 75/25 examples in training/test set, for test set;

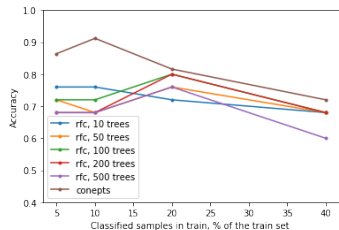
²The algorithm is adapted from Guyon and was designed to generate the “Madelon” dataset.

See details in I. Guyon, “Design of experiments for the NIPS 2003 variable selection benchmark”, 2003

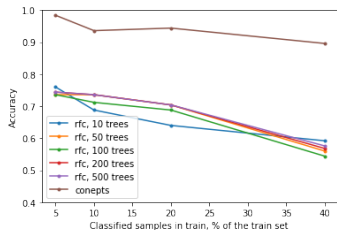
Concept-based Classifiers vs Random Forests



5 informative, 15 random



10 informative, 10 random

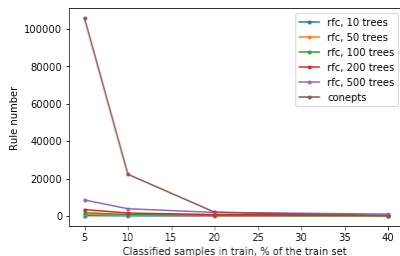


15 informative, 5 random

20 informative, 0 random

Complexity of Models

Example. 5 informative and 15 random attributes



% , min. support	number of leaves in random forest of N trees					number of concepts
	10	50	100	200	500	
5	173.6	861.6	1723.2	3416.8	8548.6	105674.0
10	78.2	386.2	772.6	1545.0	3857.8	22287.0
20	37.4	185.4	371.0	751.2	1884.0	2002.4
40	20.0	100.0	200.0	400.0	999.0	87.1

Concept-based Classifiers vs Random Forests

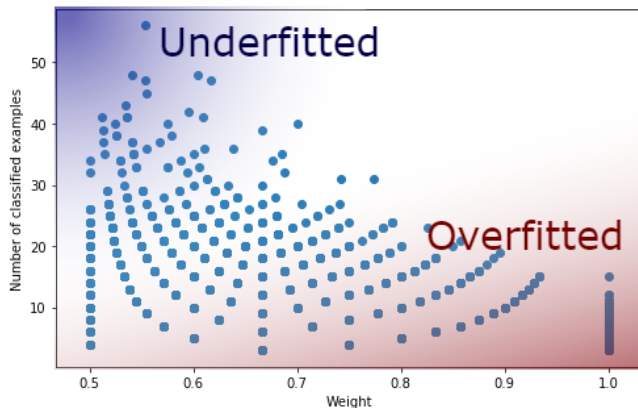
Observations

- ▶ Concept classifiers performs often better than RF / FRF.
- ▶ The number of concepts decreases exponentially with increasing threshold on the number of classified examples (on training set) by every concept.
- ▶ The number of concepts decreases exponentially with increasing accuracy of concept-based classifiers.

Can We Do Better?

Intuition: we want to select concepts which

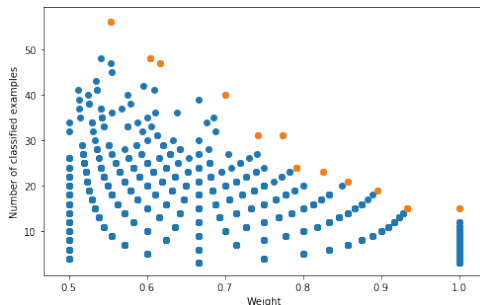
- ▶ generalize well (\equiv classify a lot of examples in training set)
- ▶ quite accurate (\equiv classify mostly examples from one class).



Concept-based Classifiers on Pareto-Optimal Concepts

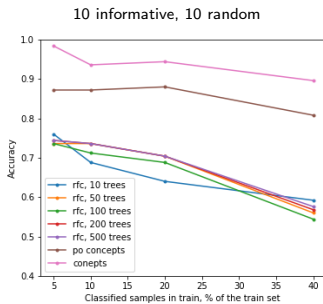
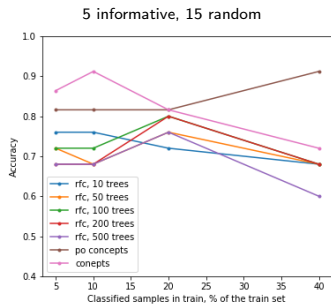
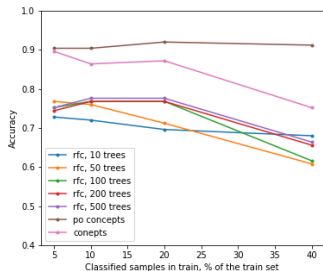
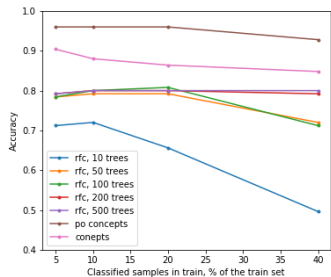
Intuition:

- ▶ select both accurate (weights are close to 1) and inaccurate ones to avoid overfitting;
- ▶ select the concepts with the maximal number of classified objects in training set to ensure good generalisation ability.



Pareto optimal concepts (exact)

Pareto Optimal Concepts. Performance



15 informative, 5 random

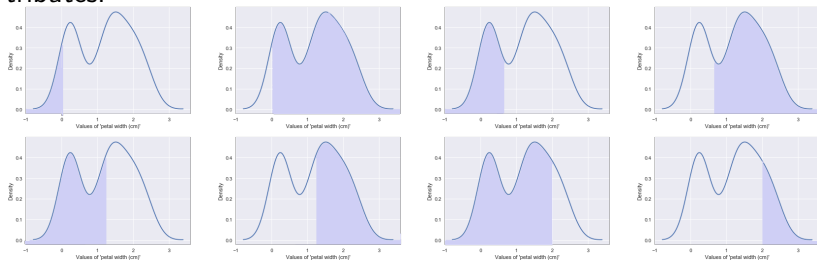
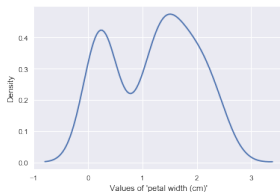
20 informative, 0 random

Efficiency

% , min. support	number of leafs in random forest of N trees					number of concepts	number of pareto concepts
	10	50	100	200	500		
5	173.6	861.6	1723.2	3416.8	8548.6	105674.0	50.0
10	78.2	386.2	772.6	1545.0	3857.8	22287.0	50.0
20	37.4	185.4	371.0	751.2	1884.0	2002.4	30.2
40	20.0	100.0	200.0	400.0	999.0	87.1	12.0

Application to Real-World Data

Numeric data: scaling: less / greater
then or equal attributes.
One threshold point – two binary at-
tributes.



Iris Dataset. Performance on Scaled Data

Threshold points are quantiles: 40%, 45%, 50%, 55%, 60% for each attribute

			Concepts	Random Forest
setosa	vs	others	0.58	0.97
versicolor	vs	others	0.76	0.97
virginicav	vs	others	0.86	0.97
versicolor	vs	virginica	0.96	0.92
setosa	vs	virginica	1.00	1.00
setosa	vs	versicolor	0.88	0.88

- ▶ Concept-based classifiers perform better on **class-balanced data**.
- ▶ Open question is the **proper choice of thresholds**.
- ▶ Application an extension of formal concepts, i.e. pattern structures.

Thank you for your attention