Prediction rule ensembles: Balancing accuracy and interpretability

Marjolein Fokkema
Leiden University
m.fokkema@fsw.leidenuniv.nl

1

Aim: Identify subgroups and/or (combinations of) predictive factors.

E.g.:
- Which children are at risk for delinquency / academic problems?
- Which variables predict academic / medical / social outcomes?
- Which (combinations of) items predict the criterion?

2

Decision trees (CART, CHAID, …)

+ Identify subgroups and variables
+ Capture non-linear associations
- Selection bias towards variables with more possible cutpoints
- Instability

E.g.:
- Which children are at risk for delinquency / academic problems?
- Which variables predict academic / medical / social outcomes?
- Which (combinations of) items predict the criterion?

3

Multiple regression (GLM, elastic net, …)

+ Identify variables
+ Capture non-linear associations
- Does not capture non-linearities
- Does not identify subgroups

E.g.:
- Which children are at risk for delinquency / academic problems?
- Which variables predict academic / medical / social outcomes?
- Which (combinations of) items predict the criterion?

4

Example: Predicting chronic depression

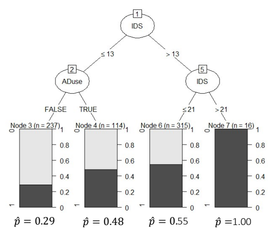
Sample: Respondents with current depressive disorder (N = 682)
Response: Depression diagnosis at two-year follow-up

20 possible predictors (baseline):
- gender, age, years of completed education
- presence of anxiety disorder(s)
- IDS (depressive symptoms)
- Receiving pharmacotherapy, psychotherapy
- BAI and FQ (anxiety symptoms)
- ….

5

Conditional inference tree

(Hothorn et al., 2006)



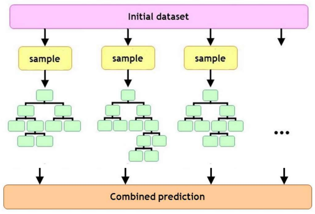
6

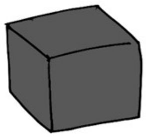
Single trees

Good: Easy to interpret and apply
Bad: Not most accurate method
Ugly: Unstable

7

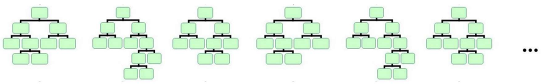
Tree ensembles
(random forest, gradient boosting, ...)





8

Tree ensembles
(random forest, gradient boosting, ...)



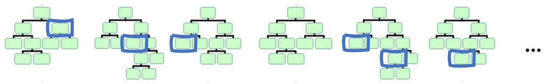
9

Prediction rule ensembling

RuleFit; Friedman & Popescu (2008)

Aim: Combine forces of trees, forests and GLM

- ❑ Keep most important rules only

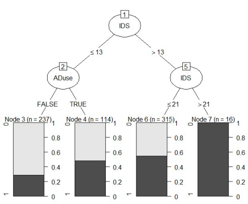


- ❑ + Linear effects of predictors
- ❑ + Penalized regression

10

From trees to rules

$r_2(\mathbf{x}) = I(IDS \leq 13)$
$r_3(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = FALSE)$
$r_4(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = TRUE)$
$\cancel{r_5(\mathbf{x}) = I(IDS > 13)}$
$r_6(\mathbf{x}) = I(IDS > 13) \cdot I(IDS \leq 21)$
$r_7(\mathbf{x}) = I(IDS > 13) \cdot I(IDS > 21)$



11

From trees to rules

$r_2(\mathbf{x}) = I(IDS \leq 13)$
$r_3(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = FALSE)$
$r_4(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = TRUE)$
$r_6(\mathbf{x}) = I(IDS > 13) \cdot I(IDS \leq 21)$
$r_7(\mathbf{x}) = I(IDS > 13) \cdot I(IDS > 21)$
...
$l_1(\mathbf{x}) = IDS$
$l_2(\mathbf{x}) = ADuse$
...

IDS	ADuse	...	r_2	r_3	r_4	r_6	r_7	...
5	FALSE	...	1	1	0	0	0	...
15	FALSE	...	0	0	0	1	0	...
18	TRUE	...	0	0	0	1	0	...
25	TRUE	...	0	0	0	0	1	...
...	...	...	...	...	...	...	...	...

Predictive model:

$$\boxed{F(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta f_m(\mathbf{x})}$$

Estimate by lasso regression:

$$\boxed{\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}}$$

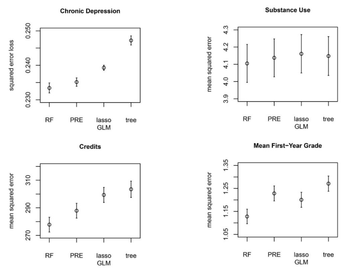
12

Example: Predicting chronic depression

Rule	Coefficient
(Intercept)	-0.218
IDS > 10 & perc_symp_2yrs > 0.263	0.246
IDS > 13 & perc_symp_2yrs > 0.362	0.152
IDS > 10 & perc_symp_2yrs > 0.328	0.140
IDS ≤ 16 & age_onset > 17	-0.083
perc_symp_2yrs > 0.260 & IDS > 9	0.016
IDS ≤ 16 & GAD = FALSE	-0.005

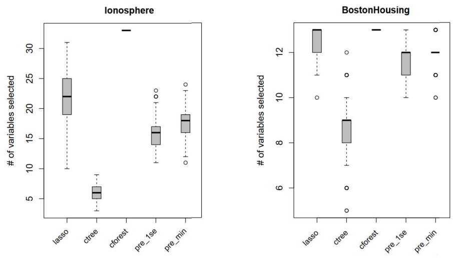
13

Predictive Accuracy
Fokkema & Strobl (2020) Psych Methods



14

Sparsity
Fokkema (2020) J of Stat Softw



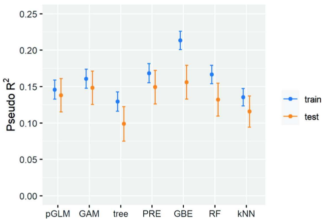
15

Predicting university major
Fokkema et al. (2022), Eur J Psych Assessment

- N = 55,593 (75% training; 25% test)
- Took psychology as a major at university: Yes (19.4%) vs. No
- Predictors: 48 items on a vocational preference scales

16

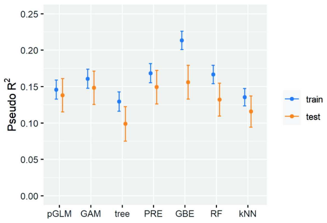
Predicting university major
Fokkema et al. (2022), Eur J Psych Assessment



(p)GLM = (penalized) logistic regression
GAM = generalized additive model with smoothing splines
PRE = prediction rule ensemble
GBE = gradient boosted tree ensemble
RF = random forest
*k*NN = *k* nearest neighbors

17

Predicting university major
Fokkema et al. (2022), Eur J Psych Assessment



(p)GLM retained 46 items
GAM retained all 48 items
tree retained 70 splits
PRE retained 41 items (134 rules)
GBE retained all 48 items (3500 trees)
RF retained all items (500 trees)
*k*NN is a black box

18

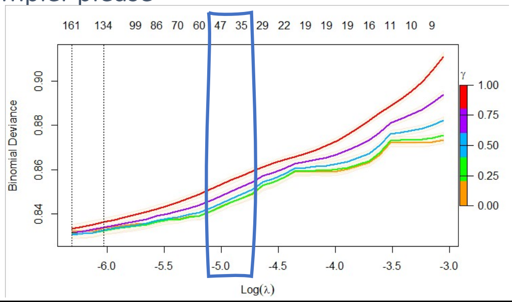
Simpler please?

- Lasso works well for discarding variables
- But overshrinks coefficients of important predictors
  - Meinshausen (2007): Relaxed lasso
- What if I want to retain a pre-specified number of rules?
  - I.e., do not optimize predictive accuracy only

- Hastie et al. (2020) describe the **glmnet** implementation of relaxed lasso
  - To use it, add relax = TRUE in the call to function cv.glmnet()
  - To use it in **pre**, add relax = TRUE in the call to function pre()

19

Simpler please



20

Interpretable ànd explainable AI

-> PRE is interpretable <u>and</u> explains black box

Problems:

- Complexity (too many rules retained by lasso)
- Lower accuracy than black box
- Instability (trees, lasso)

21

R package:

Fokkema, M. & Christoffersen, B. **pre**: Prediction Rule Ensembles:

https://github.com/marjoleinF/pre

https://cran.r-project.org/package=pre

22

Fokkema (2020). Fitting prediction rule ensembles with R package pre. *Journal of Statistical Software:* http://doi.org/10.18637/jss.v092.i12

Fokkema & Strobl (2020). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods:* http://doi.org/10.1037/met0000256

Fokkema, Iliescu, Greiff & Ziegler (2022). Machine learning and prediction in psychological assessment: Some promises and pittfalls. *Eur J of Psychological Assessment:* https://doi.org/10.1027/1015-5759/a000714

Friedman, J. & Popescu, B.E. (2008). Predictive Learning via rule ensembles. *Annals of Applied Statistics*, 2(3), 916-954.

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, *35*(4), 579-592.

23