Prediction rule ensembles (PREs)

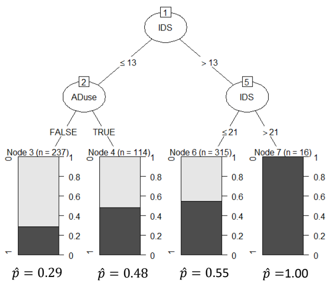
An interpretable machine-learning method

Example dataset

Penninx et al. (2011): Predict chronic depression
Sample: Respondents with current depressive disorder (N = 682)
Response: Depression diagnosis (at two-year follow-up)

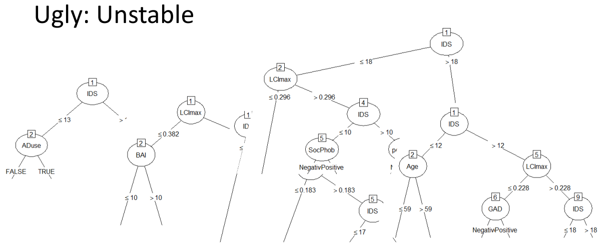
20 possible predictors (at baseline):
- Gender, age, years of completed education
- Type of depressive and/or anxiety disorder(s)
- Symptom scale scores on depression and anxiety
- Receiving pharmacotherapy, psychotherapy
- ....

Conditional inference tree (Hothorn et al., 2006)



Single trees

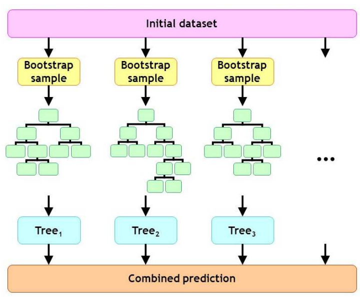
Good: Easy to interpret and apply
Bad: Not most accurate method
Ugly: Unstable



Single trees

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

Tree ensembles



Tree ensembles

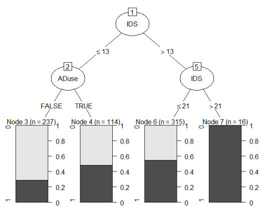
- Good: High predictive accuracy
- Bad: Difficult to interpret and apply
  - Many (complex) trees
  - Prediction requires lots of computation and information
- Prediction rule ensembles: Only keep parts that contribute most to accuracy. E.g.:
  - RuleFit (Friedman & Popescu, 2008)
  - Node Harvest (Meinshausen, 2010)
  - …

RuleFit (Friedman & Popescu, 2008)

1) Draw samples from training data
2) Fit tree on each sample
   - Classification and regression tree (CART) algorithm
   - Boosting (learning rate > 0)
3) Create initial ensemble, comprising
   - every node from every tree as a predictor (rule) and
   - every original predictor variable as a predictor
4) Select final ensemble by sparse regression on training data
   - Lasso, ridge or elastic net

From trees to rules

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



From trees to rules

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

$$\boxed{F(\mathbf{x}) = \hat{\alpha}_0 + \sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})}$$

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	…
…	…	…	…	…	…	…	…	…

RuleFit (Friedman & Popescu, 2008)

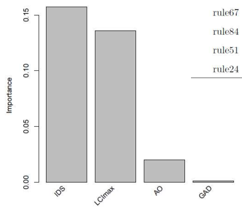
1) Draw samples from training data
2) Fit tree on each sample
   - Classification and regression tree (CART) algorithm
   - Boosting (learning rate > 0)
3) Create initial ensemble, comprising
   - every node from every tree as a predictor (rule) and
   - every original predictor variable as a predictor
4) Select final ensemble by sparse regression on training data
   - Lasso, ridge or elastic net

R package **pre** (Fokkema & Christoffersen, 2019)

1) Draw samples from training data
2) Fit tree on each sample
   - Unbiased recursive partitioning (Hothorn et al., 2006)
   - Boosting (learning rate > 0)
   - Random forest (mtry < p)
3) Create initial ensemble, comprising
   - every node from every tree as a predictor (rule) and
   - every original predictor variable as a predictor
4) Select final ensemble by sparse regression on training data
   - Lasso, ridge or elastic net

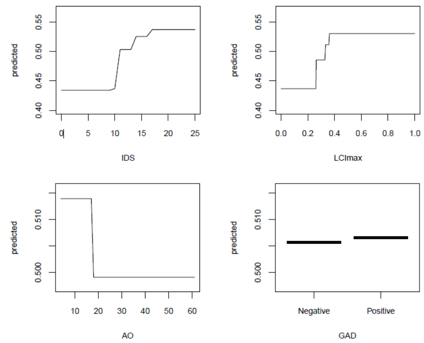
\+ ….

PRE for predicting chronic depression



Term	Coefficient	Description	SD	Importance
(Intercept)	-0.218	1		
rule3	0.246	IDS > 10 & LCImax > 0.263	0.494	0.121
rule27	0.152	IDS > 13 & LCImax > 0.362	0.477	0.073
rule67	0.140	IDS > 10 & LCImax > 0.328	0.500	0.070
rule84	-0.083	IDS ≤ 16 & AO > 17	0.489	0.041
rule51	0.016	LCImax > 0.260 & IDS > 9	0.487	0.008
rule24	-0.005	IDS ≤ 16 & GAD ∈ {Negative}	0.499	0.002

PRE for predicting chronic depression



Additional features

- (Non-) negativity constraints:
  - In order to only identify higher-risk (or lower-risk) subgroups, can enforce rules with positive (or negative) coefficients only to be selected
- Supported response variable types:
  - Binary, multinomial
  - (Multivariate) Continuous
  - Counts
  - Survival
- 'Confirmatory' rules:
  - Apply no penalty to rules or predic predictive of the response

Term	Coefficient	Description
(Intercept)	-0.218	1
rule3	0.246	IDS > 10 & LCImax > 0.263
rule27	0.152	IDS > 13 & LCImax > 0.362
rule67	0.140	IDS > 10 & LCImax > 0.328
rule84	-0.083	IDS ≤ 16 & AO > 17
rule51	0.016	LCImax > 0.260 & IDS > 9
rule24	-0.005	IDS ≤ 16 & GAD ∈ {Negative}

Example dataset 2

- Campbell et al. (2014): RCT comparing outpatient treatments for drug abuse
  - TAU vs. TES (TAU + therapeutic education system)
- 478 participants with complete data
- Response: No. of substance use days in last week of treatment
- 56 potential predictors:
  - Socio-demographic variables
  - Items measuring:
    - Quality of life
    - Coping strategies
    - Mental health problems
  - ...

- Receiving TES included as confirmatory rule

Predicting substance use

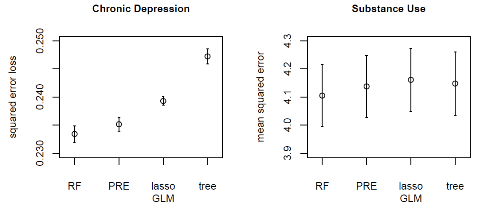
Term	Coefficient	Description
(Intercept)	0.559	1
rule20	-0.195	week1 ≤ 0 & BSNAUSE.T0 ≤ 2
trt ∈ {TES}	-0.177	trt ∈ {TES}
rule16	-0.157	week1 ≤ 2 & CSCALM.T0 > 2
rule30	-0.120	week1 ≤ 0 & BSTENSE.T0 ≤ 3
week1	0.060	0 ≤ week1 ≤ 7

Resolution

Does **pre** eliminate the bad?

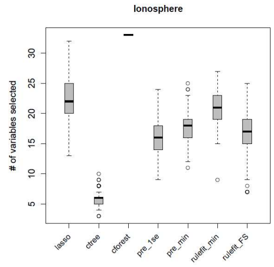
Does the good survive?

Predictive accuracy



Results based on 100 CV repeats (Fokkema & Strobl, in press)

Interpretability



Results based on 250 CV repeats (Fokkema, in press)

Contributions

- PREs balance accuracy (of tree ensembles) and interpretability (of single trees)
- Package **pre** improves on original RuleFit algorithm:
  - Selects lower number of rules and variables
  - Yields higher predictive accuracy
  - Support for
    - Several types of response variables
    - (Non-) negativity constraints
    - Confirmatory rules

Challenges

- Predictions are more stable, but the fitted model (selected rules and linear terms and their coefficients) still shows instability
  - Property inherited from (lasso) regression and decision trees
  - Not unique for these methods. E.g., Effron (2019): Prediction is easy, attribution is difficult
- Future work:
  - Dealing with missing data
  - Better (i.e., more sparse, more stable) rule and variable selection:
    - Alternatives to lasso / glmnet
    - Accounting for multilevel structures
    - …

Software and further reading

Fokkema, M. & Christoffersen, B. (2019). **pre**: Prediction Rule Ensembles. **R** package version 0.7.1 (available from CRAN).
url: https://github.com/marjoleinF/pre

Fokkema, M. (in press). Fitting prediction rule ensembles with **R** package **pre**. *Journal of Statistical Software*.
pre-print: https://arxiv.org/abs/1707.07149

Fokkema, M. & Strobl, C. (in press). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods.*
pre-print: https://arxiv.org/abs/1907.05302

m.fokkema@fsw.leidenuniv.nl

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Classification and regression trees. Wadsworth, New York.

Campbell, A. N., Nunes, E. V., Matthews, A. G., Stitzer, M., Miele, G. M., Polsky, D., et al. (2014). Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *American Journal of Psychiatry, 171* (6), 683-690.

Effron, B. (2019). *Prediction, estimation, and attribution*. Keynote at Conference in honor of Aad van der Vaart's 60th birthday, Leiden, The Netherlands. url: http://pub.math.leidenuniv.nl/~schmidthieberaj/publications/TalksAad/Efron.pdf

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

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674.

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, 4(4), 2049-2072.

Penninx, B. W. J. H., Nolen, W. A., Lamers, F., Zitman, F. G., Smit, J. H., Spinhoven, P., . . ., Beekman, A. T. F. (2011). Two-year course of depressive and anxiety disorders: Results from the Netherlands Study of Depression and Anxiety (NESDA). *Journal of Affective Disorders, 133*(1), 76-85.