Estimating out-of-sample performance for diagnostic classifications

-

Comparing resampling methods

Christian Thiele & Gerrit Hirschfeld

30 August 2016

Hochschule Osnabrück
University of Applied Sciences

The problem

- We have some data
- Develop model for classification that is optimized / fitted on our training data

How well will these models that are 'optimal' in a specific sample generalize to other samples?

- Parametric methods based on statistics (see Altman et al., 1994)
- **Resampling Methods**
- **Goal: a comparison of the estimations of a model's performance in the population made by the different resampling methods**

Solutions

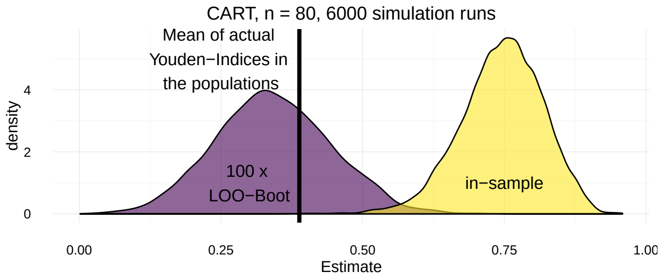
General idea of resampling methods

- Internal validation using the sample data
- We are mainly interested in the quality of future predictions
- Resampling mimicks the process of training (fitting) and testing
- **Vital:** Automate all model building and preprocessing steps and execute in every training set, otherwise potentially severe optimistic bias
- Direct estimate of the model performance or error
- Often used for model comparisons or parameter tuning

Schematic of most common resampling methods

Bias and Variance

- Depending on the sample at hand the estimate will vary
- Bias: Resampling can result in a systematically higher (or lower) estimate of the true external performance
- Variance: Even if unbiased, the estimate may vary considerably around the true external performance
- Bias/Variance Tradeoff: Resampling methods with little bias may have high variance and vice versa



Resampling methods

Training / Test split

- **Idea:** Split dataset in two sets (e.g. 75% for training and 25% for assessing performance)
- **Variants:**
  - Repetitions: Can be repeated by randomly sampling new split s n times and averaging the results

Cross validation

- **Idea:** Split into k distinct sets of size $\frac{n}{k}$. For 1 to k:
  - Fit the model using the remaining $n - \frac{n}{k}$ observations
  - Average the k test set results
- **Variants:**
  - Number of folds (k usually between 2 and 10)
  - Repetitions: Average the results of multiple CV runs

Bootstrapping

- **Idea:** Draw with replacement a random bootstrap-sample x_b with size equal to the original sample x
- **Variants:**
  - Conventional Bootstrap: Train on x_b, test on x
  - Leave-one-out bootstrap: Train on x_b, test on $x \notin x_b$
  - .632: 0.632 * conventional bootstrap + (1 - 0.632) * LOO-bootstrap
  - .632+: weight w not 0.632 but a dependent on amount of in-sample overfitting
  - optimism corrected: optimism = performance in x_b - conventional bootstrap. Subtract optimism from performance in x.

Data

A clinical data set from a children's hospital with:

- 795 observations
- Binary outcome: Migraine yes (56%) / no (44%)
- 9 independent variables (frequency of pain, auxiliary symptoms, age, etc.)

Models

Classification And Regression Trees

- Variables: All 9 variables
- Maximum tree depth: 4
- Minimum observations in a node to attempt a split: 6

Youden-based cutoff

- Variable: Number of auxiliary symptoms (6 unique values)
- Cutoff that yieds the highest sum of sensitivity and specificity.

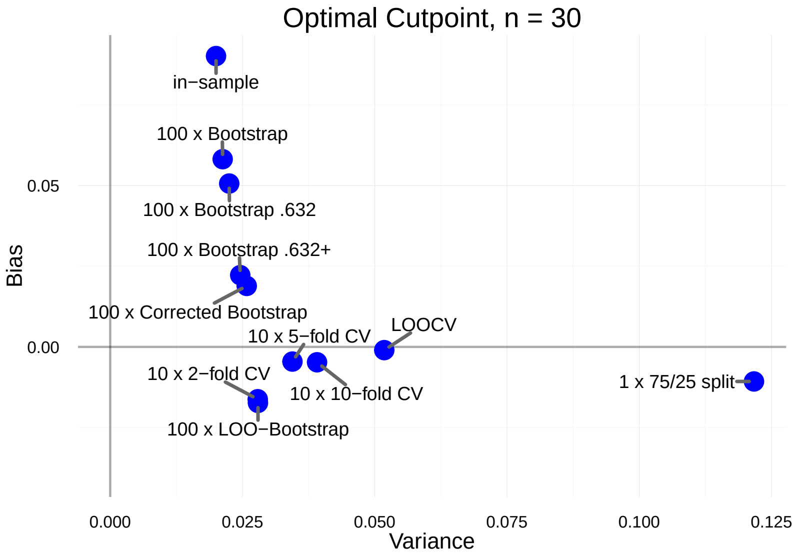
Simulation study

Nested simulation procedure

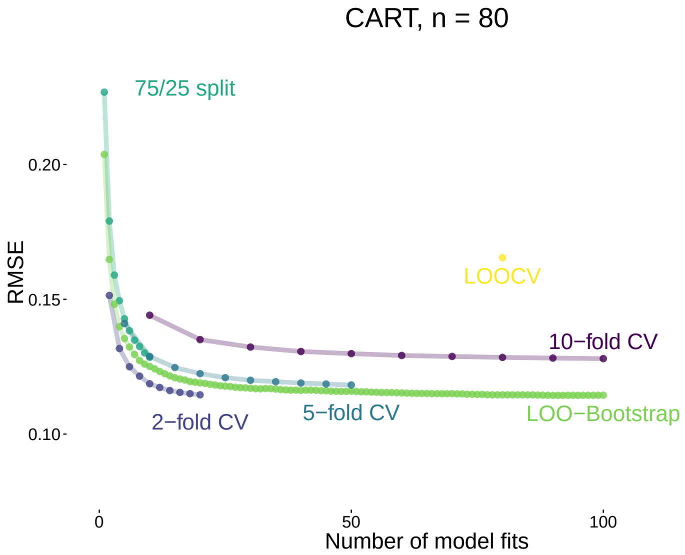
1. Draw a very small sample ($n = 30$, 45, 60 or 80) from the data set
2. Use this sample for model fitting and the validation procedure
3. Check pseudo out-of-sample performance on the rest of the data set and compare with the estimate from the internal validation
4. Compare the performance of the resampling methods in terms of bias and variance

- Drawing small samples mimics analyzing a sample from a large population
- Repeat simulation steps 6000 times (13000 in the case of training / test split)

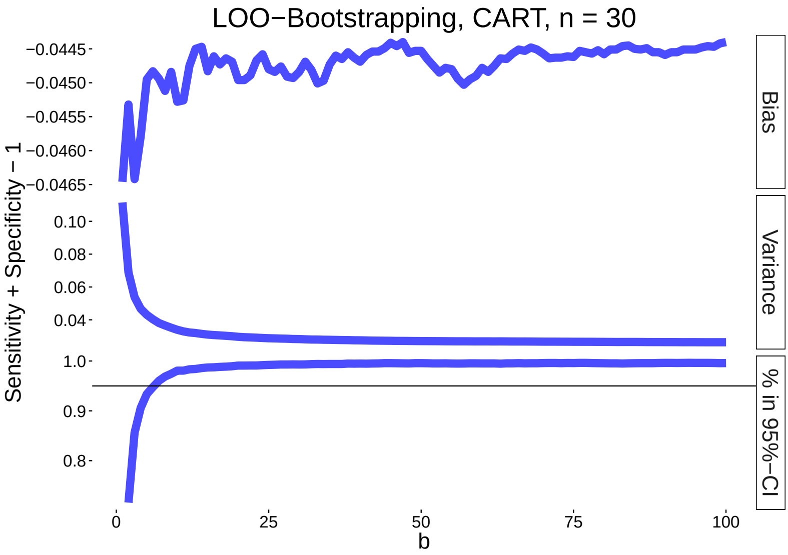
Bias and variance comparison



RMSE depending on number of model fits



Effect of repetitions in LOO-bootstrap



Main results

Bootstrapping

- Leave-one-out bootstrapping works well, low variance and good RMSE with $b \geq 50$
- Suitability of .632 variants seems to be dependent on the performance metric and/or model
- Here the optimism-corrected bootstrap has a low bias and variance with the cutpoint model (not with CART), but *optimistic* bias

Cross validation

- Pessimistically biased
- Bias lower with larger number of folds but variance rises (tradeoff!)

Leave-one-out cross validation

- Low bias, high variance

Single Training / Test split

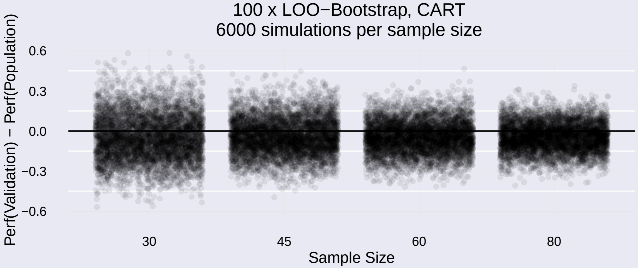
- High variance, pessimistically biased. Needs large sample size

Resampling may indentify severe forms of overestimation

- Selection of a specific model: Leave-one-out bootstrappting (low variance)
- Estimating out of sample performance: CV competitive (lower bias with $k \geq 5$)
- Differences between the validation methods most pronounced in small samples
- Confidence intervals can't be constructed using normal theory based on the resamples/splits
  - . . . unless bias and variance in that specific scenario are known, e.g. from simulations
  - Simulation computationally expensive. Example: Simulating CART in 10 x 10-fold CV 1000 times $= 10^5$ trees
- The above findings were confirmed using simulated data as well
- **External validation remains the definitive assessment of model quality**

Future research

- R-Package for running such simulations (maybe based on caret)?
- Extrapolating bias and variance e.g. from simulations with $30 <= n <= 80$ to bias and variance when $n = 795$?



Contact

- `c.thiele@hs-osnabrueck.de`
- Github.com/thie1e (Slides and markdown code)

Funding

- BMBF Indimed

Appendix

Appendix

A clinical data set with 795 observations

Patients

- 795 children (72% female; 6-24 yrs. mean = 15 years)

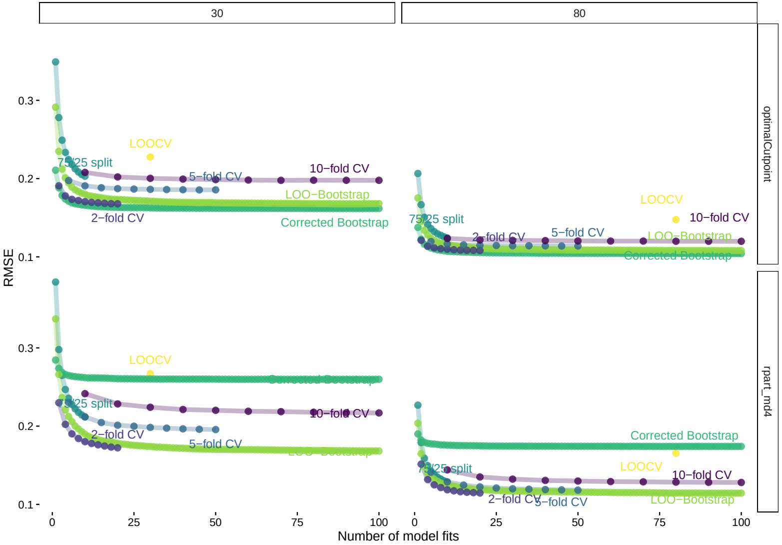
Diagnosis (after 1,5h interview with two physicians)

- Migraine or any other headache diagnosis (not migraine)
- 56% with migraine

A clinical data set with 795 observations

Variables assessed via self-report (Schroeder et al., 2010; Wager et al., 2010)

- Frequency: constant to once a year DSFKJ
- Months since first incidence
- Newly felt permanent pain (yes/no)
- Auxiliary symptoms (nausea, vomiting, sensititvity to light, loss of vision, sensitivity to sound, etc.): DSFKJ
    - Number of these symptoms
- Number of main-pain locations: DSFKJ
- Number of painful locations: DSFKJ
- Sex
- Age


30
80
optimalCutpoint
rpart_md4
RMSE
Number of model fits
0.1
0.2
0.3
0
25
50
75
100
LOOCV
75/25 split
5−fold CV
10−fold CV
LOO−Bootstrap
2−fold CV
Corrected Bootstrap

Metrics

- S Simulation runs of the validation methods $v = 1, \ldots, S$
- B data splits, cross-validation folds or bootstrap repeats: $j = 1, \ldots, B$
- iv: Internal validation estimate of the selected validation method, e.g. mean of k cross-validation folds
- g: Model performance outside of the training set (what we try to approximate)
- $Error_v$: $iv_v - g_v$
- Bias $= Mean(iv_v - g_v) = \sum_{v=1}^{S} \frac{(Error_v)}{S}$
- Variance $= Var(Error) = \frac{1}{S-1} \sum_{v=1}^{S} (Error_v - Bias)^2$
- Confidence interval: $C^{\pm} = iv \pm t_{n-1,\alpha/2} * Sd(iv)$

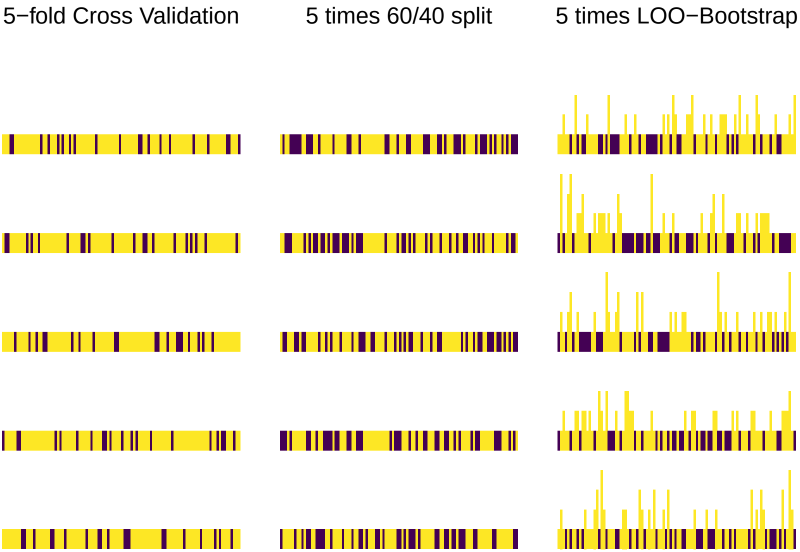
Bootstrap .632+

x: full sample
x_b: bootstrap sample
x_t: observations not in x_b
$perf$: 'Performance' in terms of the Youden-index
err: $1 - perf$
Bootstrap .632+ $= (1 - w) * perf(x_t) + w * perf(x_b))$

$$w = \frac{0.632}{1 - 0.368 * R}$$

$$R = \frac{err_{test} - err_{resub}}{err_{rand} - err_{resub}}$$

More realistic depiction of resampling



References I

Altman, D. G., Lausen, B., Sauerbrei, W., & Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst*, *86*(11), 829–835.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. Retrieved from `https://hal-ens.archives-ouvertes.fr/docs/00/40/79/06/PDF/preprintLilleArlotCelisse.pdf`

Schroeder, S., Hechler, T., Denecke, H., Müller-Busch, M., Martin, A., Menke, A., & Zernikow, B. (2010). Deutscher Schmerzfragebogen für Kinder, Jugendliche und deren Eltern (DSF-KJ). *Der Schmerz*, *24*(1), 23. `http://doi.org/10.1007/s00482-009-0864-8`

Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., & Moons, K. G. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, *56*, 441–447. Retrieved from `https://www.researchgate.net/profile/Ewout_Steyerberg/publication/10702563_Internal_and_external_validation_of_predictive_models_A_simulation_study_of_bias_and_precision_in_small_`

References II

samples/links/54abcc8a0cf2ce2df6691007.pdf

Wager, J., Tietze, A.-L., Denecke, H., Schroeder, S., Vocks, S., Kosfelder, J., . . . Hechler, T. (2010). [Pain perception of adolescents with chronic functional pain : adaptation and psychometric validation of the Pain Perception Scale (SES) by Geissner]. *Schmerz*, *24*(3), 236–250. http://doi.org/10.1007/s00482-010-0920-4