Variable Importance on Medical Images and Socio-Demographic Data

Ahmad Chamma, Denis Engemann, Bertrand Thirion

Variable Importance on Medical Images and Socio-Demographic Data

Ahmad CHAMMA
Inria, CEA, Université Paris Saclay
ahmad.chamma@inria.fr

Denis ENGEMANN
Inria, CEA, Université Paris Saclay
denis.engemann@gmail.com

Bertrand THIRION
Inria, CEA, Université Paris Saclay
bertrand.thirion@inria.fr

1 Introduction

Biomarker development targeting mental health is increasingly focusing on heterogeneous sources of data including brain images, biological samples and social data [1, 2, 3, 4]. Biobank initiatives give access to tens of thousands of brain images and unstructured social and biomedical data that can add context to the brain data [5, 6]. These large-scale datasets make it possible to predict biomedical outcomes using machine learning [7, 8, 9], shaping a novel prospective epidemiology framework. To interpret predictive models correctly and pave the way to causal assessments, it is crucial to understand how input features influence the prediction [10, 11, 12]. Over the past decades, a wide range of methods has been developed for ranking variables according to their importance in predictive models [13, 14, 15]. However, given the variety of settings (e.g. dimensionality or non-linearities, classification vs regression) it remains unclear which method provides the most accurate feature ranking for the given prediction task [16, 17]. Benchmarks have been conducted for multiple methods using simulations and empirical validation [18, 19, 20, 21], yet these efforts have been disconnected so far because of the diversity of research settings [22]. As a result, some of the most popular methods for estimating variable importance have never been systematically compared.

Here, we extend the literature by systematically comparing the most popular methods for linear and non-linear inputs in classification and regression tasks. For methods providing assessment of statistical significance, we assessed if the p-values are well calibrated. We also put performance metrics in perspective with computation time.

2 Experiments

2.1 Simulated and Real Data

The data $\boldsymbol{X}$ are generated under two scenarios: correlated and independent predictors. In the correlated scenario, we generated the data with a pre-specified correlation of 0.8. We fix the number of variables p to 50 where the number of samples to 1000. In this benchmark, we used 5 different models to generate the outcome $\boldsymbol{y}$ from $\boldsymbol{X}$:

Classification: We multiply the $\boldsymbol{n}_{\text{signal}}$ columns with $\boldsymbol{\beta}$ coefficients using the cumulative function of the standard normal distribution F. $\boldsymbol{\beta}$ has only $\boldsymbol{n}_{\text{signal}} = 20$ non-zero coefficients, the true model support. Following [20], the $\boldsymbol{\beta}$ values are drawn from the set $\{\pm 3, \pm 2, \pm 1, \pm 0.5\}$.

$$\boldsymbol{y} \sim \textbf{Binomial}(p = \mathrm{F}(\boldsymbol{X}.\boldsymbol{\beta})) \tag{1}$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Simple Regression: We rely on a linear model, where $\boldsymbol{\beta}$ is drawn as in the previous case and $\boldsymbol{\epsilon}$ is the gaussian additive noise $\sim \mathcal{N}(0, 1)$

$$\boldsymbol{y} = \boldsymbol{X} \,.\, \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

Regression with Relu: An extra ReLu function is applied to the generated data of the previous case.

$$\boldsymbol{y} = \textbf{Relu}(\boldsymbol{X} \,.\, \boldsymbol{\beta} + \boldsymbol{\epsilon}) \tag{3}$$

Interactions only model: We compute the product of each pair of variables. The corresponding resulted values are used as inputs to a linear model.

$$\boldsymbol{y} = \sum_{\substack{i,j=1 \\ i<j}}^{n_{signals}} \boldsymbol{\beta}_{i,j}.X_i X_j \tag{4}$$

The $\binom{n_{signals}}{2}$ non-zero coefficients are drawn as described previously.

Main effects with Interactions: We combine both Main and Interaction effects (Simple Regression and Interactions only model).

$$\boldsymbol{y} = \boldsymbol{X} \,.\, \boldsymbol{\beta}^{\text{main}} + \sum_{\substack{i,j=1 \\ i<j}}^{n_{signals}} \boldsymbol{\beta}^{\text{inter}}_{i,j}.X_i X_j \tag{5}$$

Real Dataset: In [9], MRI and sociodemographic data were found to provide independent information on biological aging and health in the the UK Biobank dataset. However, the importance of individual variables was not analyzed. Here, we considered the subset of 8360 samples, 2155 variables analyzed in [9]. We predicted the chronological age from MRI data, education, lifestyle, mental health and early life.

2.2 A benchmark of representative state-of-the-art methods

Marginal Effects: a univariate (generalized) linear model model is fit to explain the response from each of the variables, separately. The importance scores are then obtained from the magnitude of the coefficient (equivalent ranking would be obtained from the corresponding p-value).

Knockoffs [15]: the model fits the original features along some generated copies that are conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{X}$. The importance score is the difference between the importance of a given feature and the importance of its knockoff.

Shapley values (SHAP) [11]: SHAP being an instance method, we relied on an aggregation (averaging) of the per-sample Shapley values.

Mean Decrease Impurity (MDI) [13]: the importance scores are related to the impact one feature has on the impurity function in each of the nodes.

$\mathbf{d}_0$CRT [18]: The Conditional randomization Test is a conditional feature importance test using a random forest estimator, that also provides p-values. Distillation makes it more computationally efficient.

BART [19]: A sum-of-trees approach, where trees are constrained by a prior regularization. Fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm. Importance scores simply count the occurrence of each variable in the model.

Deep Neural Network (DNN) [21]: the importance scores are the expected value of the squared difference between the predictions using the original data and a permuted version of the variable of interest. p-values are computed by assuming normality of the importance scores.

2.3 Evaluation metrics (100 repetitions of the simulation, fit and evaluation process)

AUC score [23]: It measures how consistent the importance based variable ranking is with the binary ground truth ($n_{signals}$ predictive feature versus $p - n_{signals}$).

Type I error: Some of the methods output p-values for each of the variables, that measure the evidence against each variable being a null variable. This score checks whether the rate of low p-values is not exceeding the nominal false positive rate.

Power: This score reports the average proportion of informative variables detected (p-value < 0.05).

Computation time: The average computation time per core on 5 cores.

3 Results

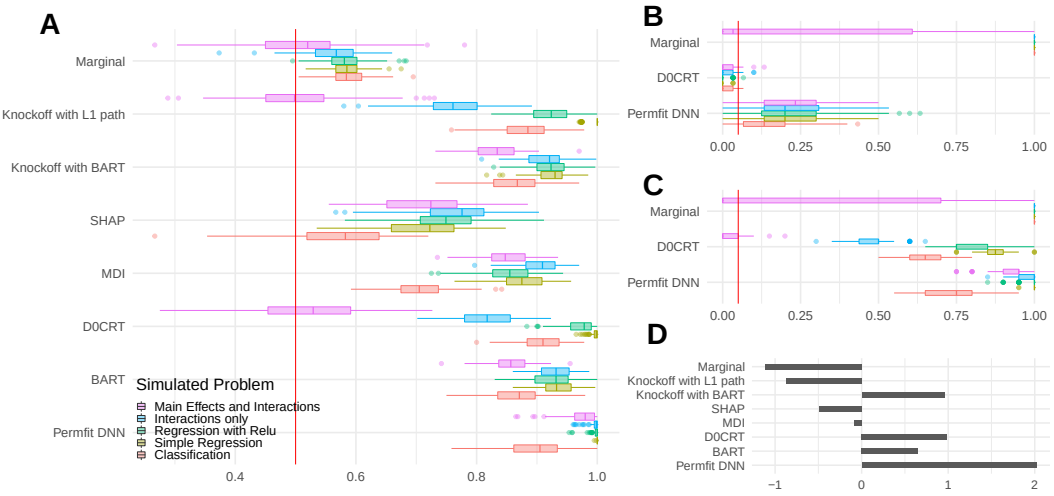

Figure 1: Experimental results on simulated data. A: AUC Score, B: Type I Error, C: Power, D: Running time in mins in log10 scale. Data were generated with correlation ($\rho = 0.8$, $n = 1000$, $p = 50$), under five scenarios.

The results for AUC score, Type I error, Power and computation time are presented in figure 1 A, B, C and D respectively, only for correlated data. Based on AUC, we observe that Marginal Effects, SHAP and Knockoff with L1 regularization perform poorly. These approaches are vulnerable to correlation. Next, d_0CRT and Knockoff with BART perform well when the model does not include interaction effects. Mean Decrease Impurity (MDI) and BART show very good performance overall. Finally, DNN outperforms all the other methods. Considering false-positive control, in the correlated setting, only d_0CRT controls type-I error. In the independent setting, Marginal and DNN also control type-I errors (not shown). The other methods do not provide p-values. Regarding power, DNN outperforms alternatives. Applying DNN and Bart on the IDPs from UKBB, Table 1, the importance scores for both methods dropped exponentially with higher decay for DNN. They showed the importance of employment status, brain's volume and income/number in household in the prediction of age (importance rankings differed for the two methods).

Table 1: Top 6 ranked variables in age prediction, for the UK BioBank dataset, $p = 2155, n = 8360$.

Features with DNN	Importance
Current employment status	4.273
Length of time at current address	1.395
Volume of grey matter (normalized)	0.912
Number in household	0.798
Volume of grey matter in Ventral Striatum (left)	0.248
Length of working week for main job	0.177

Features with BART	Importance ($\times 10^{-2}$)
Volume of grey matter (normalized)	1.815
Work/job satisfaction	1.506
Volume of peripheral cortical grey matter	1.497
Income before tax	1.32
Time employed in main current job	1.262
Number in household	1.187

4 Discussion

Deep Neural Network (DNN) is the most reliable method to select variables according to their importance. Yet, it fails to control type-1 error when variables are correlated. Moreover, the current implementation is orders of magnitude slower than the others as seen in Fig. 1 D, calling for more efficient implementations. We notice that SHAP, one of the best instance-level methods, fails to return reliable population-level importance scores. BART and MDI represent an interesting tradeoff between

computation time and reliability, but do not provide statistical guarantees. Marginal selection and knockoffs do not generalize well beyond the linear case and suffer from design correlation. Finally, d0CRT struggled with the interaction effects despite the use of Random Forests.

References

[1] A. Coravos, S. Khozin, and K. Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. 2019.

[2] J. Sieber. Integrated biomarker discovery: combining heterogeneous data. 2011.

[3] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae, R. Janardan, H. Liu, G. Alexander, and E. Reiman. Heterogeneous data fusion for alzheimer's disease study. 2008.

[4] D. Castillo-Barnes, J. Ramirez, F. Segovia, F. Martinez-Murcia, D. Salas-Gonzalez, and J. Gorriz. Robust ensemble classification methodology for i123-ioflupane spect images and multiple heterogeneous biomarkers in the diagnosis of parkinson's disease. 2018.

[5] R. Mitchell and C. Waldby. National biobanks: Clinical labor, risk production, and the creation of biovalue. 2009.

[6] N. Allen, C. Sudlow, T. Peakman, and R. Collins. Uk biobank data: Come and get it. 2014.

[7] S. Smith, Y. Fan, D. Vidaurre, F. Alfaro-Almagro, T. Nichols, and K. Miller. Estimation of brain age delta from brain imaging. 2019.

[8] L. de Nooij, M. Harris, M. Adams, T. Clarke, X. Shen, S. Cox, A. Mclntosh, and H. Whalley. Cognitive functioning and lifetime major depressive disorder in uk biobank. 2020.

[9] K. Dadi, G. Varoquaux, J. Houenou, D. Bzdok, B. Thirion, and D. Engemann. Population modeling with machine learning can enhance measures of mental health. 2020.

[10] C. Molnar. A guide for making black box models explainable. 2021.

[11] P. Biecek and T. Burzykowski. Explanatory model analysis. 2021.

[12] H. Ishwaran. Variable importance in binary regression trees and forests. 2007.

[13] G. Louppe, L. Wehenkel, A. Sutera, and P Geurts. Understanding variable importances in forests of randomized trees. 2013.

[14] C. Strobl, A. Boulseteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. 2008.

[15] E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. 2017.

[16] S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. 2018.

[17] I. Lee and Y. Shin. Machine learning for enterprises: Applications, algorithm selection, and challenges. 2019.

[18] M. Liu, E. Katsevich, L. Janson, and A. Ramdas. Conditional randomization test. 2006.

[19] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. 2010.

[20] S. Janitza, E. Celik, , and A. Boulseteix. A computationally fast variable importance test for random forests for high-dimensional data. 2015.

[21] X. Mi, B. Zou, F. Zou, and J. Hu. Permutation-based identification of important biomarkers for complex diseases via machine learning models. 2021.

[22] M. Altenmuller. When research is me-search: How researchers' motivation to pursue a topic affects laypeople's trust in science. 2021.

[23] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. 1997.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]