Hornung, Roman; **Boulesteix, Anne-Laure**

Interaction forests: identifying and exploiting interpretable quantitative and qualitative interaction effects. (English) Zbl 07512643

Comput. Stat. Data Anal. 171, Article ID 107460, 18 p. (2022)

Summary: Although interaction effects can be exploited to improve predictions and allow for valuable insights into covariate interplay, they are given limited attention in analysis. *Interaction forests* are a variant of random forests for categorical, continuous, and survival outcomes that explicitly models quantitative and qualitative interaction effects in bivariable splits performed by the trees constituting the forests. The new *effect importance measure* (EIM) associated with interaction forests allows for ranking of covariate pairs with respect to their interaction effects' importance to prediction. Using EIM, separate importance value lists for univariable effects, quantitative interaction effects, and qualitative interaction effects are obtained. In the spirit of interpretable machine learning, the bivariable split types of interaction forests target easily interpretable and communicable interaction effects. To learn about the nature of the interplay between covariates identified as interacting it is convenient to visualise their estimated bivariable influence. Functions that perform this task are provided in the R package *diversityForest*, which implements interaction forests. In a large-scale empirical study using 220 data sets, interaction forests tended to deliver better predictions than conventional random forests and competing random forest variants that use multivariable splitting. In a simulation study, EIM delivered considerably better rankings for the relevant quantitative and qualitative interaction effects than competing approaches. These results indicate that interaction forests are suitable tools for the challenging task of identifying and making use of easily interpretable and communicable interaction effects in predictive modelling.

MSC:

62-XX Statistics

Keywords:

interaction effects; random forest; feature importance; non-parametric modeling; machine learning

Software:

SHAFF; SNPInterForest; diversityForest; obliqueRF; Tunability; ranger; XGBoost; OpenML; R

Full Text: DOI

References:

[1] Basu, S.; Kumbier, K.; Brown, J. B.; Yu, B., Iterative random forests to discover predictive and stable high-order interactions, Proc. Natl. Acad. Sci. USA, 115, 8, 1943-1948 (2018) · Zbl 1416.62594

[2] Bénard, C.; Biau, G.; da Veiga, S.; Scornet, E., Interpretable random forests via rule extraction, (Banerjee, A.; Fukumizu, K., Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (2021)), 937-945

[3] Bénard, C.; Biau, G.; da Veiga, S.; Scornet, E., SHAFF: fast and consistent SHApley eFfect estimates via random Forests (2021)

[4] Bertsimas, D.; Dunn, J., Optimal classification trees, Mach. Learn., 106, 1039-1082 (2017) · Zbl 1455.68159

[5] Boulesteix, A. L.; Janitza, S.; Hapfelmeier, A.; Van Steen, K.; Strobl, C., Letter to the editor: on the term 'interaction' and related phrases in the literature on random forests, Brief. Bioinform., 16, 2, 338-345 (2015)

[6] Boulesteix, A. L.; Stierle, V.; Hapfelmeier, A., Publication bias in methodological computational research, Cancer Inform., 14, 11-19 (2015)

[7] Breiman, L., Random forests, Mach. Learn., 45, 1, 5-32 (2001) · Zbl 1007.68152

[8] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Ston, C. J., Classification and Regression Trees (1984), Wadsworth International Group: Wadsworth International Group Monterey, CA · Zbl 0541.62042

[9] Bureau, A.; Dupuis, J.; Falls, K.; Lunetta, K. L.; Hayward, B.; Keith, T. P., Identifying SNPs predictive of phenotype using random forests, Genet. Epidemiol., 28, 171-182 (2005)

[10] Chen, T.; Guestrin, C., XGBoost: a scalable tree boosting system, (Krishnapuram, B.; Shah, M., Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)), 785-794

[11] Chen, Z.; Zhang, W., Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight, PLoS Comput. Biol., 9, 3, Article e1002956 pp. (2013)

[12] Couronné, R.; Probst, P.; Boulesteix, A. L., Random forest versus logistic regression: a large-scale benchmark experiment, BMC Bioinform., 19, 270 (2018)

[13] Dazard, J. E.; Ishwaran, H.; Mehlotra, R.; Weinberg, A.; Zimmerman, P., Ensemble survival tree models to reveal pairwise interactions of variables with time-to-events outcomes in low-dimensional setting, Stat. Appl. Genet. Mol. Biol., 17, 1, Article 20170038 pp. (2018) · Zbl 1398.92010

[14] Du, J.; Linero, A., Interaction detection with Bayesian decision tree ensembles, (Chaudhuri, K.; Sugiyama, M., The 22nd International Conference on Artificial Intelligence and Statistics (2019)), 108-117

[15] Gashler, M.; Giraud-Carrier, C.; Martinez, T., Decision tree ensemble: small heterogeneous is better than large homogeneous, (Wani, M. A.; Chen, X. W.; Casasent, D.; Kurgan, L. A.; Hu, T.; Hafeez, K., Seventh International Conference on Machine Learning and Applications (2008)), 900-905

[16] Geurts, P.; Ernst, D.; Wehenkel, L., Extremely randomized trees, Mach. Learn., 63, 1, 3-42 (2006) · Zbl 1110.68124

[17] Hapfelmeier, A.; Hothorn, T.; Ulm, K.; Strobl, C., A new variable importance measure for random forests with missing data, Stat. Comput., 24, 21-34 (2014) · Zbl 1325.62011

[18] Hornung, R., Diversity forests: using split sampling to enable innovative complex split procedures in random forests, SN Comput. Sci., 3, 2, 1 (2022)

[19] Ishwaran, H., Variable importance in binary regression trees and forests, Electron. J. Stat., 1, 519-537 (2007) · Zbl 1320.62158

[20] Janitza, S.; Celik, E.; Boulesteix, A. L., A computationally fast variable importance test for random forests for high-dimensional data, Adv. Data Anal. Classif., 12, 885-915 (2018) · Zbl 1416.62606

[21] Jiang, R.; Tang, W.; Wu, X.; Fu, W., A random forest approach to the detection of epistatic interactions in case-control studies, BMC Bioinform., 10, Suppl. 1, S65 (2009)

[22] Kelly, C.; Okada, K., Variable interaction measures with random forest classifiers, (Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI) (2012)), 154-157

[23] Kim, H.; Loh, W. Y., Classification trees with unbiased multiway splits, J. Am. Stat. Assoc., 96, 454, 589-604 (2001)

[24] Li, J.; Malley, J. D.; Andrew, A. S.; Karagas, M. R.; Moore, J. H., Detecting gene-gene interactions using a permutation-based random forest method, BioData Min., 9, 14 (2016)

[25] Loh, W. Y., Regression trees with unbiased variable selection and interaction detection, Stat. Sin., 12, 361-386 (2002) · Zbl 0998.62042

[26] Menze, B. H.; Kelm, B. M.; Splitthoff, D. N.; Koethe, U.; Hamprecht, F. A., On oblique random forests, (Gunopulos, D.; Hofmann, T.; Malerba, D.; Vazirgiannis, M., European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2011)), 453-469

[27] Molnar, C.; Casalicchio, G.; Bischl, B., Interpretable machine learning - a brief history, state-of-the-art and challenges, (Koprinska, I.; etal., ECML PKDD 2020 Workshops. ECML PKDD 2020. ECML PKDD 2020 Workshops. ECML PKDD 2020, Communications in Computer and Information Science, vol. 1323 (2020)), 417-431

[28] Ng, V. W.; Breiman, L., Bivariate variable selection for classification problem (2005), Department of Statistics, University of California: Department of Statistics, University of California Berkeley, CA, Technical report 692

[29] Peto, R., Statistical aspects of cancer trials, (Halnam, K. E., Treatment of Cancer (1982), Chapman \& Hall: Chapman \& Hall London), 867-871

[30] Poterie, A.; Dupuy, J. F.; Monbet, V.; Rouvière, L., Classification tree algorithm for grouped variables, Comput. Stat., 34, 1613-1648 (2019) · Zbl 1505.62323

[31] Probst, P.; Boulesteix, A. L.; Bischl, B., Tunability: importance of hyperparameters of machine learning algorithms, J. Mach. Learn. Res., 20, 53, 1-32 (2019) · Zbl 1485.68226

[32] Rainforth, T.; Wood, F., Canonical correlation forests (2015)

[33] Rodríguez, J. J.; Kuncheva, L. I.; Alonso, C. J., Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell., 28, 10, 1619-1630 (2006)

[34] Seibold, H.; Bernau, C.; Boulesteix, A. L.; De Bin, R., On the choice and influence of the number of boosting steps for high-dimensional linear Cox-models, Comput. Stat., 33, 1195-1215 (2018) · Zbl 1417.65056

[35] Shapley, L. S., A value for n-person games, (Kuhn, H. W.; Tucker, A. W., Contributions to the Theory of Games II. Contributions to the Theory of Games II, Annals of Mathematics Studies, vol. 28 (1953), Princeton University Press: Princeton University Press Princeton), 307-317 · Zbl 0050.14404

[36] Sorokina, D.; Caruana, R.; Riedewald, M., Additive groves of regression trees, (Kok, J. N.; Koronacki, J.; Mantaras, R. L.; Matwin, S.; Mladenič, D.; Skowron, A., Proceedings of the 18th European Conference on Machine Learning (2007)), 323-334

[37] Sorokina, D.; Caruana, R.; Riedewald, M.; Fink, D., Detecting statistical interactions with additive groves of trees, (Cohen, W.; McCallum, A. K.; Roweis, S. T., Proceedings of the 25th International Conference on Machine Learning (2008)), 1000-1007

[38] Strobl, C.; Boulesteix, A. L.; Zeileis, A.; Hothorn, T., Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinform., 8, 25 (2007)

[39] Vanschoren, J.; van Rijn, J. N.; Bischl, B.; Torgo, L., OpenML: networked science in machine learning, ACM SIGKDD Explor. Newsl., 15, 2, 49-60 (2013)

[40] Wright, M. N.; König, I. R., Splitting on categorical predictors in random forests, PeerJ, 7, Article e6339 pp. (2019)

[41] Wright, M. N.; Ziegler, A., ranger: A fast implementation of random forests for high dimensional data in C++ and R, J. Stat. Softw., 77, 1-17 (2017)

[42] Wright, M. N.; Ziegler, A.; König, I. R., Do little interactions get lost in dark random forests?, BMC Bioinform., 17, 145 (2016)

[43] Yoshida, M.; Koike, A., SNPInterForest: a new method for detecting epistatic interactions, BMC Bioinform., 12, 469 (2011)

[44] Zhou, M.; Dai, M.; Yao, Y.; Liu, J.; Yang, C.; Peng, H., BOLT-SSI: a statistical approach to screening interaction effects for ultra-high dimensional data (2019)