Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

© 2022 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220418

Robust Random Forest-Based All-Relevant Feature Ranks for Trustworthy AI

Bastian PFEIFER[a,1], Andreas HOLZINGER[a,b] and Michael G. SCHIMEK[a]
[a] *Institute for Medical Informatics Statistics and Documentations,*
Medical University of Graz, Austria
[b] *Alberta Machine Intelligence Institute, Canada*

Abstract. Feature selection is a fundamental challenge in machine learning. For instance in bioinformatics, it is essential when one wishes to detect biomarkers. Tree-based methods are predominantly used for this purpose. In this paper, we study the stability of the feature selection methods BORUTA, VITA, and RRF (regularized random forest). In particular, we investigate the feature ranking instability of the associated stochastic algorithms. For stabilization of the feature ranks, we propose to compute consensus values from multiple feature selection runs, applying rank aggregation techniques. Our results show that these consolidated features are more accurate and robust, which helps to make practical machine learning applications more trustworthy.

Keywords. Feature Selection, Random Forest, Rank Aggregation, Trustworthy AI

1. Motivation

Feature selection is an important preprocessing step in many machine learning applications and has long been a fundamental challenge. In the biomedical field, feature selection is commonly used for data-driven biomarker discovery. Most common feature selection methods are based on the random forest (RF) classifier because it provides an interpretable mechanism for computing feature importance. Very important aspects of trustworthy AI are robustness and explainability [1]. A robust feature selector should (1) report on a constant set of relevant features when executed on exactly the same data set multiple times (stability), and (2) select the most relevant features for the modeling process of interest. In this work we analyzed the robustness of three feature selection methods which are widely used for data-driven biomarker discovery, namely the BORUTA algorithm [2], the VITA algorithm [3], and the RRF algorithm [4]. In essence, we are proposing to run the above mentioned algorithms multiple times and to consolidate the observed rank variations of importance scores through rank aggregation techniques. In simulation experiments on synthetic data we could show, that the RF consensus feature ranks obtained via rank aggregation can substantially improve the selection of the most important and best performing features.

[1] Corresponding Author, Bastian Pfeifer, Institute for Medical Informatics Statistics and Documentation, Medical University of Graz, Austria; E-mail: bastian.pfeifer@medunigraz.at.

2. Results

For evaluation we used the Madelon data sets from the the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/madelon). Madelon is synthetic and contains data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube randomly labeled +1 or -1. Madelon includes 20 relevant and 480 irrelevant features. The performance of the feature selection algorithms was analyzed while subsequentially down-sampling the data (15%, 25%, 50%, and 75%), leading to a stepwise reduction of the relevant signals. The feature selection algorithms were executed 20 times on exactly the same down-sampled data set. The rank aggregation techniques Borda (l2norm, mean, geometric mean, and median; [5]) and alternatively RRA [6] were applied for the consensus calculations. To evaluate the stability of the feature selectors *coverage* was used, defined as *Rs/Rt*, where *Rs* is the number of relevant features which are successfully selected and *Rt* is the total number of relevant features present in the data. The analysis of the Madelon data set indicates that VITA as well as BORUTA can be improved by consolidated consensus scores. When only 25% of the data are randomly sampled, the overall performance and robustness decreases. A detailed summary of this evaluation can be obtained from Table 1, where we also varied the number of trees within the RF. We observe that for BORUTA and VITA, the consolidated consensus ranks have a much higher coverage compared to the median coverage of the 20 feature selection runs. This observation indicates that the consensus calculation may have a higher impact on robustness than an increased number of trees. Even for a number of trees as high as 1000, there is a notable difference in performance between the worst (min coverage) and the best (max coverage) run. Furthermore, we identify BORUTA as the most accurate algorithm, especially when the number of trees is low and the signal is weak. However, it is computationally more demanding than VITA. Compared to BORUTA and VITA, RRF does not benefit much from the consensus calculations.

Table 1. Madelon data set. Coverage of the consensus ranks with varying numbers of trees.

FS Method	n.trees	min/median/max	l2norm	mean	geomean	median	RRA
BORUTA	100	0.32/0.50/0.74	0.79	0.79	0.84	0.84	0.79
	500	0.58/0.68/0.79	0.74	0.74	0.79	0.74	0.74
	1000	0.53/0.68/0.79	0.74	0.74	0.74	0.74	0.74
VITA	100	0.26/0.39/0.53	0.58	0.63	0.63	0.63	0.58
	500	0.42/0.58/0.68	0.53	0.63	0.63	0.58	0.63
	1000	0.52/0.68/0.79	0.74	0.74	0.74	0.74	0.74
RRF	100	0.37/0.47/0.58	0.26	0.37	0.47	0.47	0.47
	500	0.42/0.47/0.58	0.42	0.47	0.47	0.47	0.42
	1000	0.42/0.47/0.53	0.47	0.47	0.47	0.47	0.47

References

[1] Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. *Journal of Machine Learning Research*. 2017: 18(1): 6345-6398.

[2] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J of Stat Soft*. 2010:36(11): 1–13.

[3] Janitza S, Celik E, Boulesteix AL. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*. 2018: 12(4): 885–915.

[4] Deng H, Runger G. Gene selection with guided regularized random forest. *Pattern Recognition*. 2013: 46(12): 3483-3489.

[5] Schimek MG, et al. Topklists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat App in Gen Mol Bio*. 2015:14:311-6.

[6] Kolde R, Laur S, Adler P, and Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*. 2012: 28(4): 573–580.