Journal of Data Science xx (xx), 1–6 DOI: xx.xxxx/xxxxxxxxx
July 2020

Interaction Selection and Prediction Performance in High-Dimensional Data: A Comparative Study of Statistical and Tree-Based Methods Supplementary Materials

Chinedu Jude Nzekwe *[1], Seongtae Kim[1], and Sayed Mostafa[1]

[1] *Department of Mathematics and Statistics, North Carolina Agricultural and Technical State University, Greensboro, NC, 27411, USA*

This supplementary material contains (1) the description of the Random Intersection Trees Algorithm and (2) additional simulation results for the imbalanced classification response case.

1 The Random Intersection Trees Algorithm

Algorithm 1 Random Intersection Trees (Shah and Meinshausen, 2014)

Input: $\left\{(\mathcal{I}_i, \mathcal{Z}_i); \mathcal{I}_i \subseteq \{1, \ldots, p\}, \mathcal{Z}_i \in \{0,1\}\right\}_{i=1}^{n}, \mathcal{C} \in \{0,1\}$
Tuning Parameters: (D, M, n_{child})

for *tree* $m \leftarrow 1$ to M **do**
Let m be a tree of depth D, with each node j in levels $0, \ldots, D-1$ having n_{child} children, and denote the parent of node j as $pa(j)$. Let J be the total number of nodes in the tree, and index the nodes such that for every parent-child pair, larger indices are assigned to the child than the parent. For each node, $j = 1, \ldots, J$, let i_j be a uniform sample from the set of class $\mathcal{C}$ observations $\{i : \mathcal{Z}_i = \mathcal{C}\}$
    Set $\mathcal{S}_1 = \mathcal{I}_{i_1}$

    for $j = 2$ to J **do**
        $\mathcal{S}_j \leftarrow \mathcal{I}_{i_j} \cap \mathcal{S}_{pa(j)}$
    end for
    return $\mathcal{S}_m = \mathcal{S}_j : depth(j) = D$
end for
Output: $\mathcal{S} = \cup_{m=1}^{M} \mathcal{S}_m$

2 Simulation Results for the Imbalanced Response Classification Case

Figure 1 displays the results for the imbalanced response case for the classification linear hierarchical model. While iRF shows better prediction on the training sets, it performs poorly on the testing sets.

*Author. Email: cjnzekwe@aggies.ncat.edu

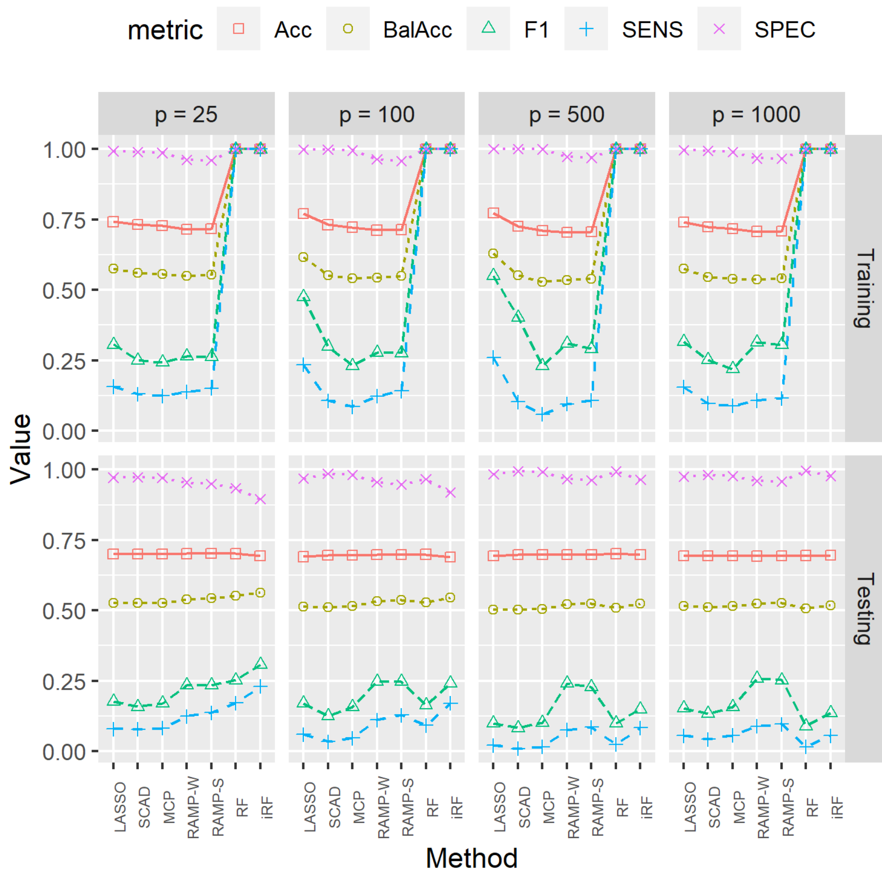

Figure 1: Classification performance metrics for seven algorithms under an imbalanced linear hierarchical model.

Table 1: Interaction selection accuracy of six classification methods under the linear hierarchical model with an imbalanced response.

	LASSO	SCAD	MCP	RAMP	RF	iRF
p	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
25	0.565 (0.348)	0.645 (0.242)	0.485 (0.233)	0.520 (0.301)	0.773 (0.263)	0.555 (0.227)
100	0.385 (0.334)	0.450 (0.273)	0.333 (0.216)	0.465 (0.284)	0.747 (0.269)	NA (NA)
500	0.203 (0.251)	0.243 (0.232)	0.170 (0.174)	0.465 (0.284)	0.600 (0.292)	NA (NA)

RAMP used here denotes RAMP-weak rule. RF shows better coverage or capturing of interaction terms but starts declining as p increases. RAMP shows stability in its selection of interaction terms for a classification linear hierarchical model with an imbalanced response case.

Table 1 compares the interaction selection coverage for six classification methods with imbalanced response cases. when $p = 25$ and $p = 100$, RF has a mean accuracy coverage of $0.773, 0.747$ respectively. For $p = 500$ RF mean accuracy 0.600 for a linear hierarchical model as shown in Table 1.

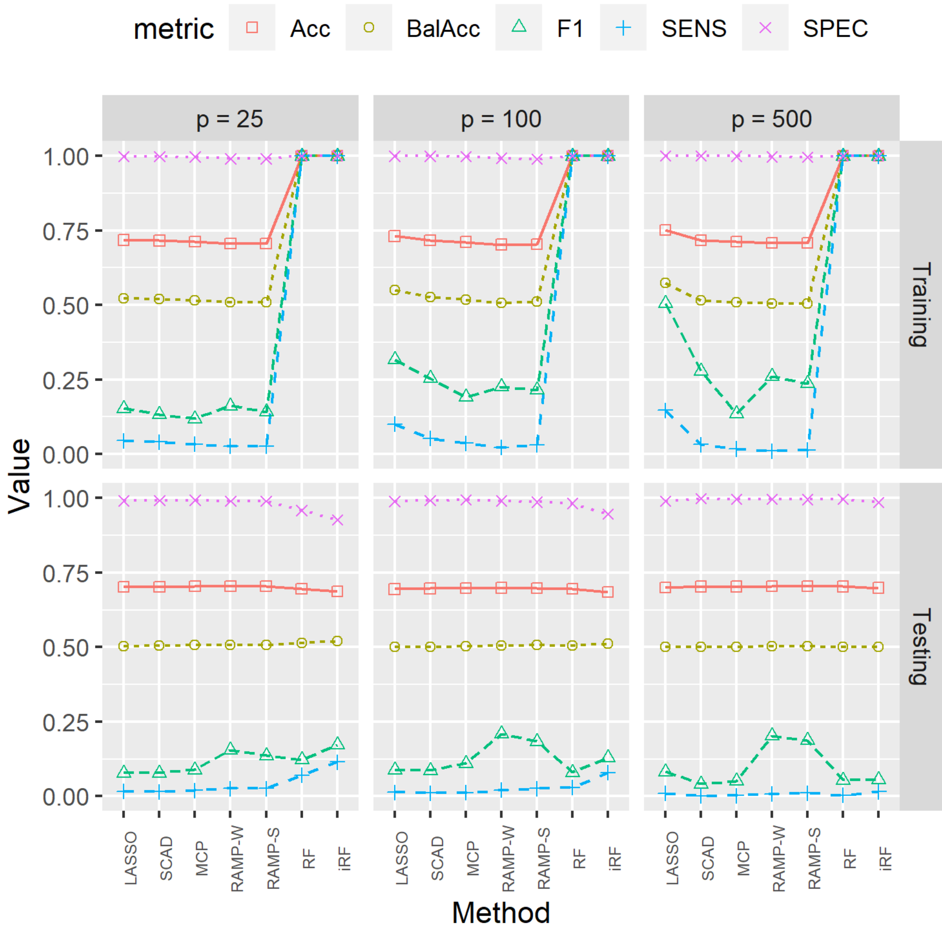

Figure 2: Classification performance metrics for seven algorithms under an imbalanced linear non-hierarchical model.

Table 2: Interaction selection accuracy of six classification methods under the linear non-hierarchical model with an imbalanced response.

	LASSO	SCAD	MCP	RAMP	RF	iRF
p	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
25	0.360 (0.420)	0.510 (0.389)	0.380 (0.390)	0.365 (0.244)	0.660 (0.476)	0.555 (0.347)
100	0.240 (0.365)	0.290 (0.342)	0.230 (0.321)	0.340 (0.234)	0.660 (0.476)	0.777 (0.427)
500	0.100 (0.213)	0.090 (0.193)	0.080 (0.197)	0.332 (0.147)	0.350 (0.479)	NA (NA)

RAMP used here denotes RAMP-weak rule. The tree-based method (RF and iRF) shows better coverage or capture of interaction terms but declines as p increases. RAMP shows stability in its selection of interaction terms for a classification linear non-hierarchical model with an imbalance response case.

For a linear non-hierarchical model, RF and iRF show the highest mean accuracy coverage of (0.660) and (0.777), for $p = 25$ and $p = 100$, respectively. However, when $p = 500$, RF has the best coverage of 0.350, while the RAMP algorithm shows stability in its capturing of interaction regardless of the size of p. The penalty-driven algorithms LASSO, SCAD, and MCP show a significant decrease in mean accuracy, with MCP having the highest mean accuracy see

Table 2. It is important to note that while iRF shows high mean accuracy at lower p values, its performance drops as p increases, particularly in the nonlinear non-hierarchical model. The RF method performs relatively better at higher p values across different models.

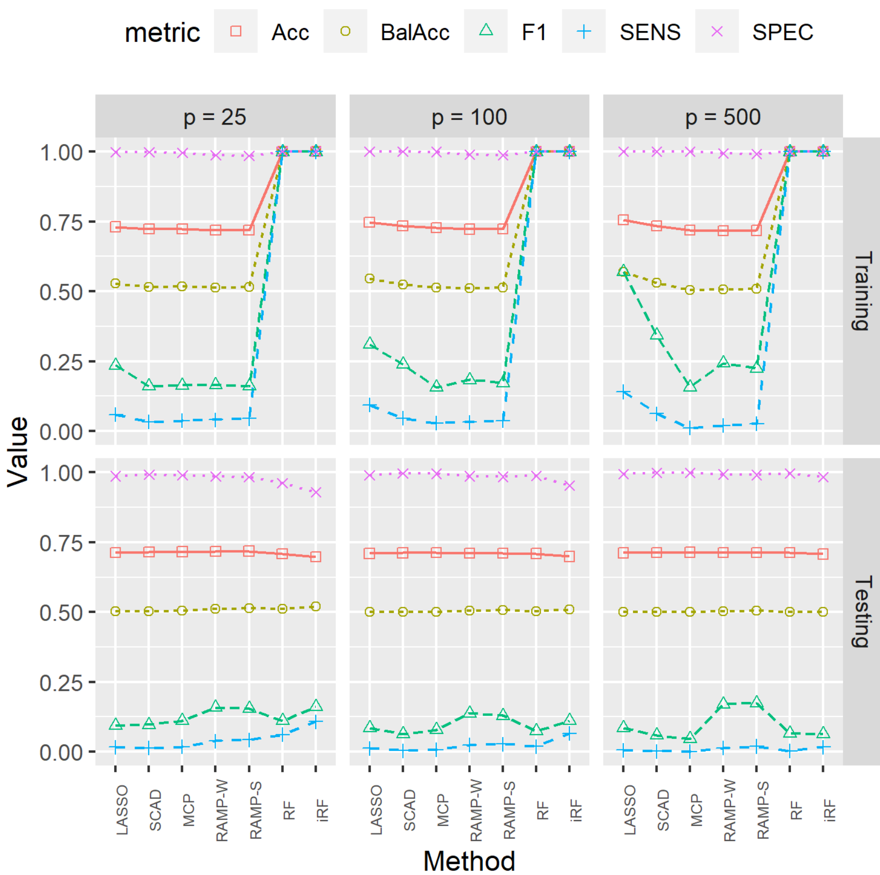

Figure 3: Classification performance metrics for seven algorithms under an imbalanced nonlinear non-hierarchical model.

Table 3: Interaction selection accuracy of six classification methods under the nonlinear model with an imbalanced response.

	LASSO	SCAD	MCP	RAMP	RF	iRF
p	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
25	0.046 (0.142)	0.060 (0.152)	0.036 (0.115)	0.040 (0.108)	0.333 (0.236)	0.598 (0.275)
100	0.003 (0.033)	0.006 (0.046)	0.003 (0.033)	0.000 (0.000)	0.300 (0.186)	0.072 (0.165)
500	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.188 (0.191)	NA (NA)

Classification performance metrics for seven algorithms under a balanced nonlinear non-hierarchical model.

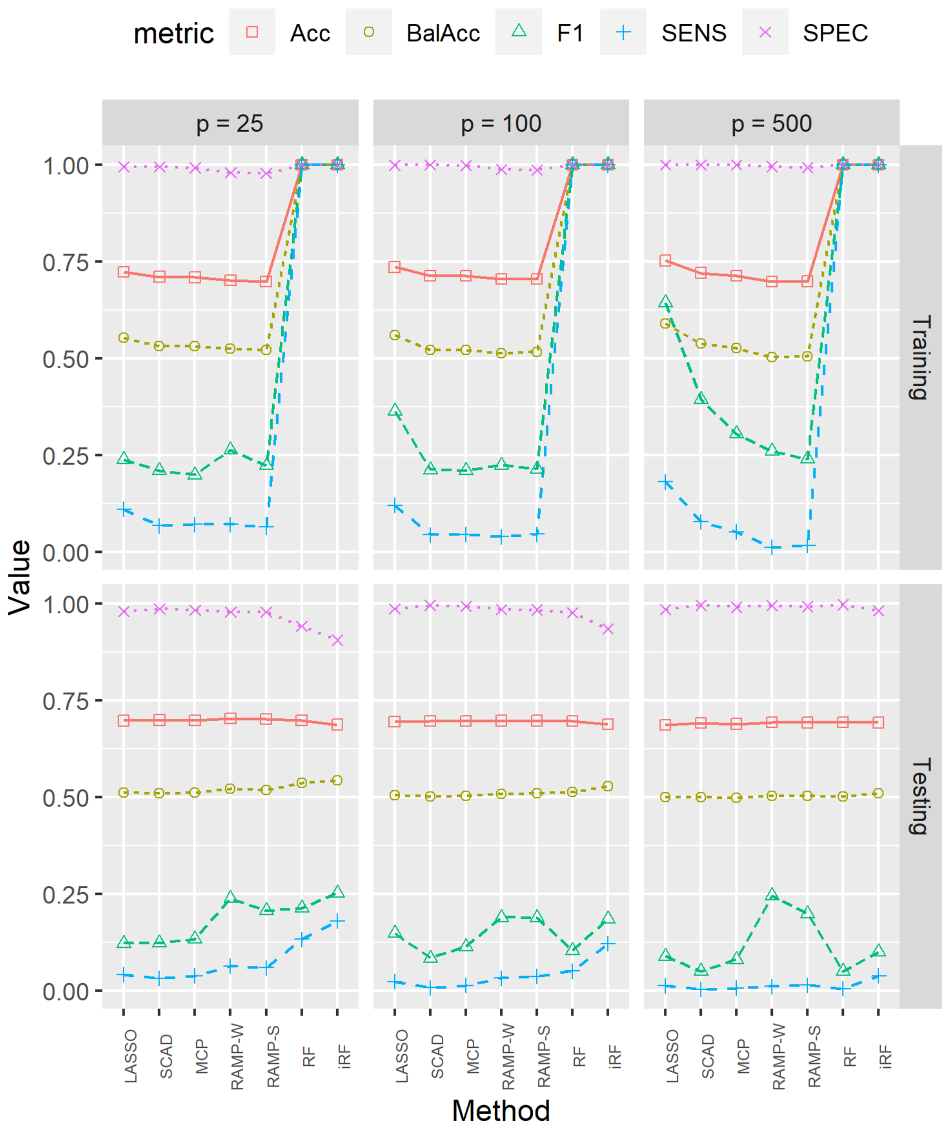

Figure 4: Classification performance metrics for seven algorithms under an imbalanced nonlinear hierarchical model.

Table 4: Interaction selection accuracy of six classification methods under the nonlinear hierarchical model with an imbalanced response.

	LASSO	SCAD	MCP	RAMP	RF	iRF
p	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
25	0.473 (0.376)	0.483 (0.358)	0.340 (0.267)	0.453 (0.330)	0.390 (0.222)	0.638 (0.252)
100	0.220 (0.311)	0.290 (0.298)	0.196 (0.255)	0.320 (0.275)	0.420 (0.245)	0.377 (0.440)
500	0.056 (0.150)	0.100 (0.214)	0.050 (0.137)	0.398 (0.211)	0.257 (0.217)	NA (NA)

RAMP used here denotes RAMP-weak rule. iRF shows better coverage or capturing of interaction terms when compared to RAMP for a classification nonlinear hierarchical model imbalanced response case.

References

Shah RD, Meinshausen N (2014). Random intersection trees. *The Journal of Machine Learning Research*, 15(1): 629–654.