Feature Selection and Imaging-Genetics Predictions Using a Sparse, Extremely Randomized Forest Regressor

Albert Montillo, Shantanu Sharma, and Marcel Prastawa
GE Global Research, Niskayuna, NY 12309

Aims: We propose a sparse extension of the extremely randomized forest (ERF) [2] nonlinear regressor by embedding it in a model reduction framework providing it with sparsity to reduce model complexity and reduced variance. The method enjoys few tunable parameters and is readily scalable to large data through parallelization. We demonstrate the utility of the method on two cases entailing joint modeling of genetic and image features with cognitive scores in Alzheimer's disease. In the first case ($\sim 10^3$) genetic SNP features are combined with the trimmed mean summary statistic of voxel shrinkage in 38 cortical and subcortical structures upon nonlinear registration to a reference brain. In the second case the SNP features are combined with quartile summary statistics of the shrinkage in a subset of 17 structures. In both the method identifies clinically relevant features by assigning feature importance scores and the final model using only relevant features achieves high prediction accuracy.

Method: We construct a model reduction framework consisting of a hierarchical cross-validation in which each fold of an outer k-fold cross-validation contains a complete q-fold inner cross-validation. The outer divides the data into train and test sets allowing for model evaluation, while the inner divides each outer train set into new q-fold train and validation sets. Similar to recursive feature elimination [3], only features whose importance is greater than the mean importance is retained for the next iteration until the validation error no longer diminishes. However for the ERF, OOB predictions [1] are unavailable therefore we use mean decrease in node impurity to compute features importance. Additionally, each run of the inner and outer cross-validation folds are repeated n=10 and m=4 times respectively with random training data shuffling to reduce variance. In the inner cross-validation feature importances are averaged across repetitions, while the outer computes optimal feature set size from votes cast by the inner cross-validation.

Results and Conclusions: We applied the proposed approach to a subset of the ADNI [5] imaging genetics data containing 30 normal and 18 AD subjects. For genomic features we normalized the $< R, \theta >$ tuples from 427 SNPs associated with AD (i.e. whose p-value $\leq 10^{-3}$) [4] that are contained in the ADNI2 GWAS panel and have resolved R and θ values. To form image features we computed the log Jacobian of the mapping between the subject's T1 MRI and a reference template. Our first imaging-genetic dataset combining our genomic features with imaging features computed as the trimmed mean (10%) of the voxel Jacobians in 38 cortical and subcortical regions defined as part of the Freesurfer atlas. Our second dataset combines the genomic features with summary quartile (Q1, Q2, Q3) measures of the Jacobian distribution of a subset of 17 structures.

We applied our method to predict AVLT [1] for both datasets. Similar RMSE prediction errors were are achieved (Fig. 1), though the best anatomical region identification occurred using quartile measures. The anatomical regions assigned high importance are shown in Fig. 2 and Fig. 3. Importances for the individual SNP R and Θ components are shown in Fig. 4 and Fig. 5. The results from our approach show promising capabilities for sparse feature selection and prediction, We look forward to applying it to additional datatsets and extending its capabilities.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 36(1):3–42, 2006.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, Mar. 2002.

[4] D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, J. S. Pahwa, V. Moskvina, K. Dowzell, A. Williams, et al. Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nature genetics*, 41(10):1088–1093, 2009.

[5] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, et al. The Alzheimer's Disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

[1] Auditory Verbal Learning Test

Features	Trimmed mean	Quartiles
Imaging-Genetics	4.86	4.47
Imaging	4.49	4.46
Genetics	4.56	4.53

Figure 1: Regression performance measured as root mean square error using 4-fold cross-validation. In terms of RMSE, both trimmed mean and quartiles give similarly good performance.

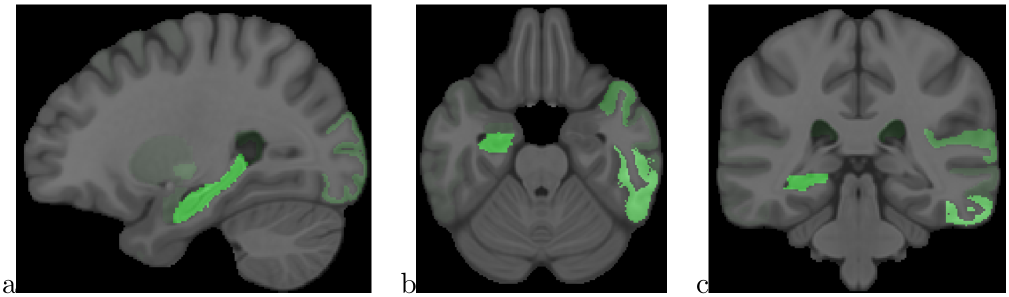

Figure 2: Automatically assigned region importances in green for AVLT prediction on ICBM template using quartile Jacobian measures. (a) Sagittal highlights hippocampus in center, while axial, coronal views (b,c) highlight hippocampus, inferior and superior temporal regions. (asymmetry from training on different structures per hemisphere.)

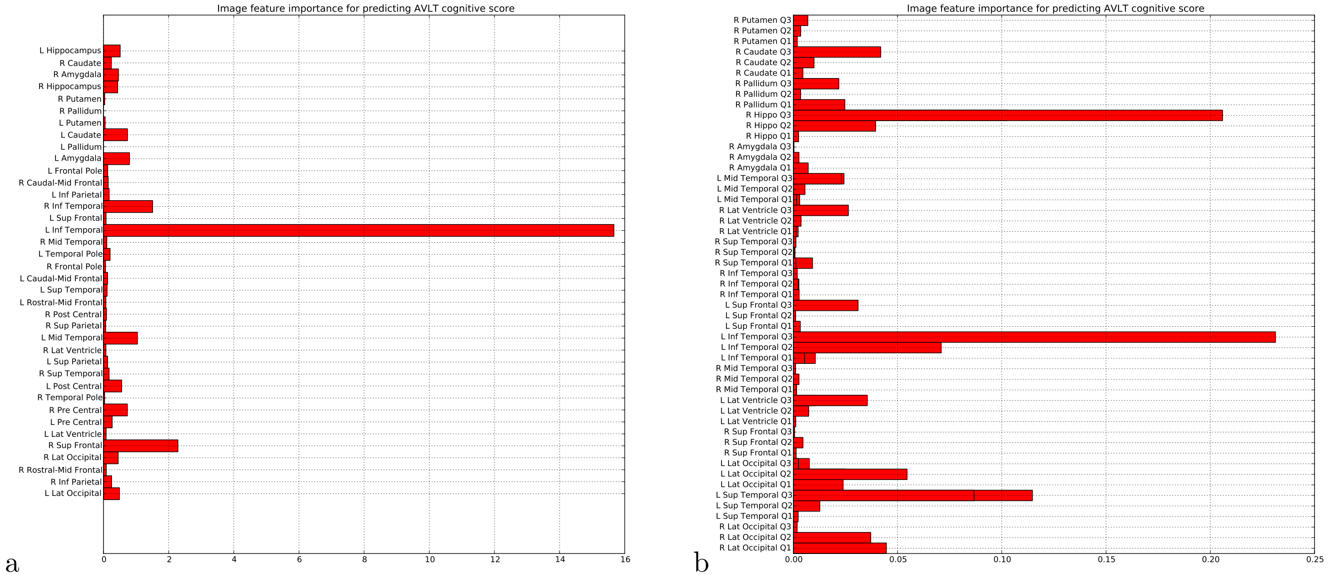

Figure 3: Importances for the imaging regions. Using trimmed mean of log Jacobian (a) yields only temporal region with high importance while using quartile measures (b) assigns high importance to hippocampus and temporal regions.

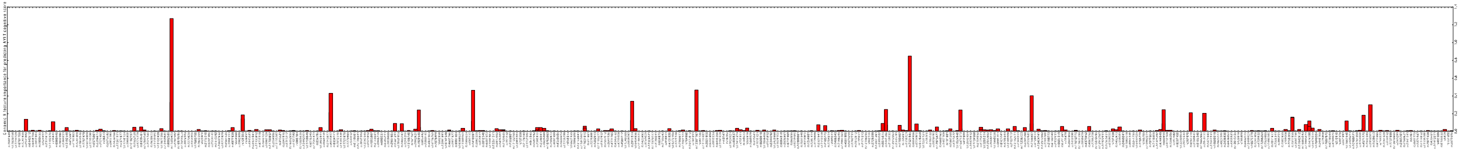

Figure 4: SNP importances for the genomic R measure.

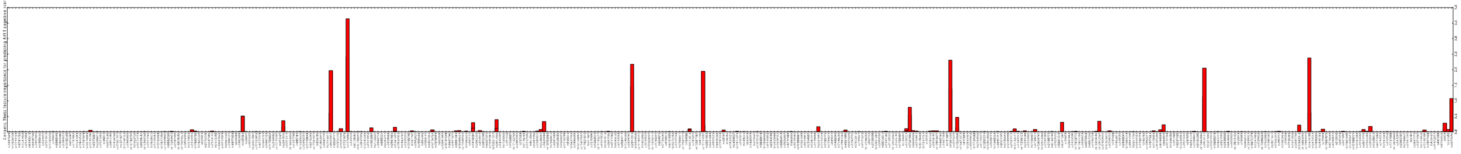

Figure 5: SNP importances for the genomic Θ measure.