Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2023;9(1):20-29

Original Article

Random-Splitting Random Forest with Multiple Mixed-Data Covariates

Mohammad Fayaz[1*], Alireza Abadi[2,3], Soheila Khodakarim[4]

[1]ECO College of Insurance, Allameh Tabataba'i University, Tehran, Iran.
[2]Department of Community Medicine, Faculty of medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[3]Social Determinants of Health Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[4]Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran.

ARTICLE INFO

Received 17.10.2022
Revised 12.12.2022
Accepted 22.01.2023
Published 15.03.2023

Key words:
Bagging;
Functional data;
Random forest;
Random splitting;
Statistical learning

ABSTRACT

Introduction: The bagging (BG) and random forest (RF) are famous supervised statistical learning methods based on the classification and regression trees. The BG and RF can deal with different types of responses such as categorical, continuous, etc. There are curves, time series, functional data, or observations that are related to each other based on their domain in many statistical applications. The RF methods are extended to some cases for functional data as covariates or responses in many pieces of literature. Among them, random-splitting is used to summarize the functional data to the multiple related summary statistics such as average, etc.
Methods: This research article extends this method and introduces the mixed data BG (MD-BG) and RF (MD-RF) algorithm for multiple functional and non-functional, or mixed and hybrid data, covariates and it calculates the variable importance plot (VIP) for each covariate.
Results: The main differences between MD-BG and MD-RF are in choosing the covariates that in the first, all covariates remain in the model but the second uses a random sample of covariates. The MD-RF helps to unmask the most important parts of functional covariates and the most important non-functional covariates.
Conclusion: We apply our methods on the two datasets of DTI and Tecator and compare their performances for continuous and categorical responses with developed R package ("RSRF") in the GitHub.

Introduction

The bagging (BG) and random forest (RF) are tree-based methods that reduce the variance of predicted models by averaging from bootstrap samples and produce the variable importance indices.[1, 2] Different approaches were developed to tackle functional data in the trees and random forests:[3] introduce two methods in the function-on-scalar (FOS) tree- regression which constructed on the spline basis with penalty term and first few functional principal component (FPCA) scores,[4] developed a R package containing methods for classification of scalar-on-function (SOF) random forest, 5 developed group variable importance for

*.Corresponding Author: mohammad.fayaz.89@gmail.com

SOF random forest based on the wavelet basis functions,[6] developed bagging FOS regression trees for density classification,[7] produced function-on-function (FOF) random forest,[8] discuss the asymptotic characteristic of the RF and their implementation for the functional setting,[9] recently introduced Fréchet trees and random forests which can handle functional data,[1] indicated that the peaks in the mass spectrum were used for data reduction and they are important landmark for constructing the Bromine tree,[10] introduce random-splitting (RS) of functional covariate on its domain and average the values between them and import the summary statistics in the RF for SOF settings,[11] extends previous method, RS, to the multiple functional covariates.

In this research, we extended the RS approach[10] to multiple functional and nonfunctional, or mixed data, covariates[12] in two ways. Firstly, considering all the covariates in the model, we converted all functional covariates into summary statistics such as means of random-splitting of the domains and all nonfunctional covariates, then placed them in BG. We called this method mix data bagging (MD-BG). The reason for these names is that in the regular bagging methods, all variables enter into the model without any sampling of covariates. Secondly, we selected a random sample of covariates, consisting of functional and nonfunctional covariates. Regarding a functional covariate, we summarized it with the random-splitting approach and placed the nonfunctional covariate in the model exactly. In each iteration of the RF, we faced a random sample of covariates. The final model was built based on these models called mix data random forest (MD-RF). The supplementary includes further analysis and the R codes.

Method

Multiple Mixed Covariates

The SOF algorithm of the MD-BG and MD-RF is summarized in Table 1. The functional covariates split randomly and the observation among them are converted into the multiple summary statistics and then the nonfunctional covariates import into the BG and RF. The main difference between MD-BG and MD-RF is that in the MD-BG, we choose all the covariates while in the MD-RF, we only chose a random sample of covariates in each tree.

The two datasets considered in this study as comprised of a DTI dataset and classification problem, and a tecator dataset and prediction problem. Regarding the first dataset, healthy and multiple sclerosis (MS) patients is a dichotomous outcome variable. The functional covariate is diffusion tensor imaging (DTI) and two other nonfunctional covariates are gender and number of visits to the clinic. The dataset was obtained from the refund R package (Goldsmith et al. 2016)[13] and the overall accuracy (ACC), sensitivity (SEN), and specificity (SPE) of the model were computed. Regarding the second dataset, the outcome variable is the percentage of fat content in a piece of meat and the functional covariate is the wavelength range 850–1050 nm by the near-infrared (NIR); also, nonfunctional covariates are the percentage of water and protein (Ferraty and Vieu 2006).[14] The dataset was obtained from the $\{i, i = 1, \ldots, I\}$ fda.usc R package (Bande et al. 2020),[15] and the mean square prediction error (MSPE) and correlation (COR) were compared with each other. The descriptive statistics for both datasets are available in the appendix section A0.

Table 1. The SOF Algorithm of MD-BG and MD-RF.

Algorithm SOF Algorithm of MD-BG and MD-RF

Steps	Description
Iterations	A. For *i*=1 to I *(Number of iterations)*
Data Preparation	1. For *k*=1 to K (Number of functional covariates) a. Split the curve $x_k(t)$, $\{t; t \in [t_l, t_U]\}$ randomly i. Choose a distribution: Exponential, Normal or Uniform. ii. Choose a statistics: Mean, Median, Range, Min, Max, Quantiles iii. For w=1 to *W* until $t + r^* > t_U$ 1. Generate a random variable *r** from chosen distribution. a. Split domain from t to $t + r^*$. b. Calculate the statistics $x_k(t), \ldots, x_k(t + r^*)$ i. If method is disjoint ,Replace $t = t + r^*$. ii. If method is overlapped ,Replace $t = t + r^* - \varepsilon$. b. Convert $x_k(t)$ to the $\overline{x}_{k1}, \ldots, \overline{x}_{kW}$ 2. Consider all, *J*, non-functional covariates $z_1, \ldots, z_J$.
	3. For b=1 to B *(Number of Trees of Forest)* a. Draw a bootstrap sample h^* of size N (Sample Size) from the training data. b. Grow a random-forest tree T_b with the bootstrap data, by recursively repeating the following the steps: i. If method is bagging (MD-BG): 1. Import all functional covariates variables $\overline{x}_{11}, \ldots, \overline{x}_{1W}, \ldots, \overline{x}_{K1}, \ldots, \overline{x}_{KW}$ 2. Import all, *J*, non-functional covariates *z1*,…,*zJ*, 3. Import the response vector y. ii. If method is random forest (MD-RF): 1. Import random functional covariates, $0 \le q_f \le K$, variables from $\overline{x}_{11}, \ldots, \overline{x}_{1W}, \ldots, \overline{x}_{K1}, \ldots, \overline{x}_{KW}$, 2. Import random samples, $0 \le q_{nf} \le J$, from non-functional covariates $z_1, \ldots, z_J$, 3. Import the response vector y. iii. Pick the best variable/split-point among them. iv. Split the node into two daughter nodes. c. Output the ensemble trees $[\{T_b\}_1^B]_i$ d. To make the prediction at a new point. i. Regression: $[\hat{f}_{rf}^B(x)]_i = \frac{1}{B}\sum_{b=1}^{B}[T_b(x)]_i$ ii. Classification: $\hat{C}_b(x)$ be the class prediction of the *bth* random-forest tree of iteration. Then $[\hat{C}_{rf}^B(x)]_i =$ [majority vote $\hat{C}_b(x)\}_1^B]_i$.
	B. To make assessment: 1. Categorical: Average over Accuracy$_i$, Sensitivty$_i$ and Specificity$_i$ $\{i, i = 1, \ldots, I\}$ 2. Continuous: Average over MSPE$_i$, Correlation$_i$ $\{i, i = 1, \ldots, I\}$.

The input covariates have two scenarios: the first is only functional covariates and the second is all covariates. The number of covariates is p and $m \approx \sqrt{p}$ for MD-BG and MD-RF, respectively (Hastie, Tibshirani, and Friedman 2009). The MD-BG and MD-RF are compared on two datasets with BG and RF. (randomForest R package (RColorBrewer and Liaw 2018)). Functional classification using ML algorithms, classif. randomForest function (FRF) from fda.usc R package.(Bande et al. 2020) The number of covariates is p and $m \approx \sqrt{p}$ for MD-BG and MD-RF, respectively.[2] The MD-BG and MD-RF are compared on two datasets with BG and RF. (randomForest R package[16]). Functional classification using ML algorithms, classif.randomForest function (FRF) from fda.usc R package.[15] is studied only for the DTI dataset.

We randomly split 70% of the data for training and 30% for testing. The number of trees in each model is 1,000 and we repeat each random forest 1,000 times. The overall ACC, SEN, and SPE for categorical outcomes and MSPE and COR for continuous outcomes in each iteration were saved. The mean, the first, and the third quartile are demonstrated in tables 2 and 3.

We write R functions with many options for summary statistics such as average (AVG); minimum (MIN); maximum (MAX); standard deviation (SD); median (MED); first and third Quartile (Q1 and Q2); disjoint and overlapping intervals as suggested in by (Möller, Tutz, and Gertheiss 2016);10; and exponential, normal and uniform distributions for r^* and ε. The running time for MD-BG and MD-RF is less than 15 min in a regular laptop with CORE-i5 CPU and 6-GB RAM for 1,000 iterations. The following R package ("RSRF") is developed and it has a vignette for further examples. With the following codes, you can install and use in it in the R: library(devtools) install_ github("mohammad-fayaz/RSRF")

The Variable Importance Plot

The variable importance plot (VIP) is a common methodology to show the importance of input covariates in the BG and RF. It calculates based on the response variable, for continuous data, indices are, type 1: mean decrease in MSE and type 2: mean decrease in node impurity, for categorical data, indices are, type 1: mean decrease in accuracy and type 2: mean decrease in node impurity. (Gareth et al. 2013; Hastie, Tibshirani, and Friedman 2009).[2, 17]

The one kind of the VIP for random-splitting was produced based on the residual and they show that the new VIP is smoother than regular VIP and therefore it enhances the interpretation of the plot. (Möller, Tutz, and Gertheiss 2016).10. There are two limitations for this plot: 1) they plot the VIP based on the average increase in the classification error rate and 2) it only works with the single functional or multiple non-functional covariates. To tackle these challenges, we propose the following VIP. The type 1 and 2 indices for each covariate, functional and non-functional, are stored for each tree of BG or RF. In the functional setting for each covariate, the random split values (minimum and maximum of functional data domain) in each tree are stored, the type 1 and 2 variable importance (VI) indices for each summary statistics of each split are calculated.

Therefore, for each curve, horizontal lines are indicating the values of type 1 or 2 of variable importance values for each split. It is for one

tree of the MD or RF. With growing many trees, these values are computed and stored for all trees on all values in the domain. Finally, the mean, median, first and third quantile of type 1 and 2 variable importance is calculated and plotted for each data points in the domain for functional covariates and their values for nonfunctional covariates. The benefit of this method are: the result is a smooth curve for each functional covariate that is easy to interpret and all observations in each functional covariate are considered together to produce the plot.

The main advantage of VIP for MD-RF over MD-BG is in incorporating the sample of covariates than all of them. Therefore, the estimated VIP doesn't mask the important features of each covariate.

Results

Multiple Mixed Covariates

The comparison between the models is presented in tables 2 and 3. The ACC and SEN of the four methods are the same with a negligible difference, but SPE in the BG and MD-BG have larger values than other methods in both A (Functional Covariate) and B (Mixed Covariate) part of table 2. There is a significant increase, about two times higher values in the SPE for part B of MD-BG and BG than part A. It indicates that the mixed covariates increase the model performance in both regular and random-splitting BG and RF. Similarly, in table 3, the MD-BG with only a Functional covariate has 72.131 (64.128, 78.981) and 70.859 (57.344, 81.605) MSPE, while these values

Table 2. The mean (First Quantile – Third Quantile) of the Accuracy, Sensitivity and Specificity in the 1,000 iterations

	Algorithm	Accuracy (ACC)		Sensitivity (SEN)		Specificity (SPE)	
		Train	Test	Train	Test	Train	Test
Functional (A)	BG	0.891 (0.882,0.901)	0.892 (0.876,0.912)	0.97 (0.966,0.974)	0.971 (0.96,0.981)	0.26 (0.2,0.312)	0.275 (0.188,0.357)
	RF	0.891 (0.882,0.901)	0.891 (0.876,0.912)	0.976 (0.97,0.979)	0.977 (0.969,0.99)	0.211 (0.148,0.273)	0.223 (0.143,0.286)
	MD-BG	0.887 (0.878,0.897)	0.889 (0.867,0.903)	0.97 (0.965,0.975)	0.97 (0.96,0.981)	0.231 (0.176,0.286)	0.244 (0.154,0.333)
	F-RF	0.891 (0.878,0.897)	0.887 (0.867,0.903)	0.973 (0.969,0.979)	0.975 (0.961,0.99)	0.193 (0.133,0.25)	0.203 (0.125,0.273)
Mixed (B)	BG	0.907 (0.901,0.916)	0.909 (0.894,0.923)	0.97 (0.965,0.978)	0.972 (0.96,0.99)	0.403 (0.346,0.464)	0.421 (0.312,0.5)
	RF	0.902 (0.894,0.909)	0.9 (0.885,0.92)	0.981 (0.978,0.987)	0.981 (0.971,0.99)	0.266 (0.207,0.333)	0.272 (0.188,0.333)
	MD-BG	0.904 (0.894,0.916)	0.907 (0.894,0.92)	0.959 (0.952,0.966)	0.961 (0.949,0.979)	0.462 (0.405,0.517)	0.488 (0.375,0.6)
	MD-RF	0.894 (0.882,0.905)	0.893 (0.876,0.912)	0.977 (0.962,1)	0.978 (0.961,1)	0.223 (0,0.394)	0.23 (0,0.4)

Models are Functional: Patient_Status~ DTI, Mixed: Patient_Status~ DTI+Number of Visit+Gender Algorithms: (BG): Bagging – (RF): Random Forest – (MD-BG): Mixed Bagging – (MD-RF): Mixed Random Forest - (F-RF): Functional Random Forest (classif.randomForest from fda.usc) – Total iterations: 1,000

Table 3. The mean (First Quantile – Third Quantile) of Mean Squared Prediction Error (MSPE) and Correlation (COR)

Covariates	Algorithm	MSPE		COR	
		Train	Test	Train	Test
Functional	BG	62.288 (59.094,65.621)	60.583 (51.749,68.362)	0.613 (0.589,0.638)	0.62 (0.582,0.677)
	RF	64.613 (61.335,68)	64.286 (54.586,72.394)	0.598 (0.575,0.623)	0.598 (0.548,0.658)
	MD-BG	72.131 (64.128,78.981)	70.859 (57.344,81.605)	0.552 (0.506,0.606)	0.556 (0.489,0.638)
Mixed	BG	2.529 (2.321,2.736)	2.494 (2.011,2.929)	0.984 (0.983,0.986)	0.984 (0.981,0.988)
	RF	23.492 (22.358,24.63)	22.819 (18.289,26.585)	0.854 (0.846,0.863)	0.858 (0.839,0.88)
	MD-BG	2.543 (2.347,2.747)	2.491 (1.993,2.916)	0.984 (0.983,0.986)	0.984 (0.981,0.988)
	MD-RF	42.604 (4.591,63.584)	41.633 (4.806,64.699)	0.736 (0.604,0.971)	0.735 (0.589,0.969)

Models are Functional: %Fat~ NIR, Mixed: %FAT~ NIR+%Water+%Protein Algorithms: (BG): Bagging – (RF): Random Forest – (MD-BG): Mixed Bagging – (MD-RF): Mixed Random Forest - Total iterations: 1,000

fall to 2.543 (2.347, 2.747) and 2.491 (1.993, 2.916) MSPE in the mixed covariates settings for train and test, respectively.

A considerable increase in their COR was observed from 0.552 and 0.556 to 0.984 and 0.984 in the train and test, respectively. It shows the importance of nonfunctional covariates in the prediction (Further analysis is provided in Appendix A.1). Tables 1 and 2 in appendix A.1. show the performance of the models by adding each input variable and compare the results.

The Variable Importance Plot

Figure 1 indicates the VIP for MD-RF based on the two types of variable importance for categorical response in the DTI dataset. It shows that in the functional covariates, the values between 65 to 75 for CCA (left-panels) and values between 20 and 30 for RCST (middle-panels) have the highest important part of their curves, and in the non-functional covariates (right-panels) the number of visits covariate have the highest value for variable importance measures. At a glance, the number of visits is the most important covariate and some parts of the CCA.

In figure 2, we compare the VIP for MD-BG and MD-RF for continuous output in the tecator dataset. Figure 2_1 for MD-BG indicates that non-functional covariates (second-row), Water, have the highest VI measures and in the functional covariate, NIT has the highest VI between 35 and 45, but their values are very low compared to the non-function covariates.

On the other hand, the VIP for MD-RF is shown in figure 2_2. The main difference is that in the MD-RF, in each tree, there is randomly chosen two from three covariates, $p = 3, \sqrt{3} \approx 2$, (one functional and two non-functional) remain in the model and the VI measures are calculated based on them. It helps to unmask and reveal the most important parts of the functional covariate

and the most important variables in the non-functional covariates. For instance, between 35 to 45 and 90 to 100 are the most important parts of the NIT curve. It is different from the figure 2_1 result for the functional part but is the same for the non-functional part.

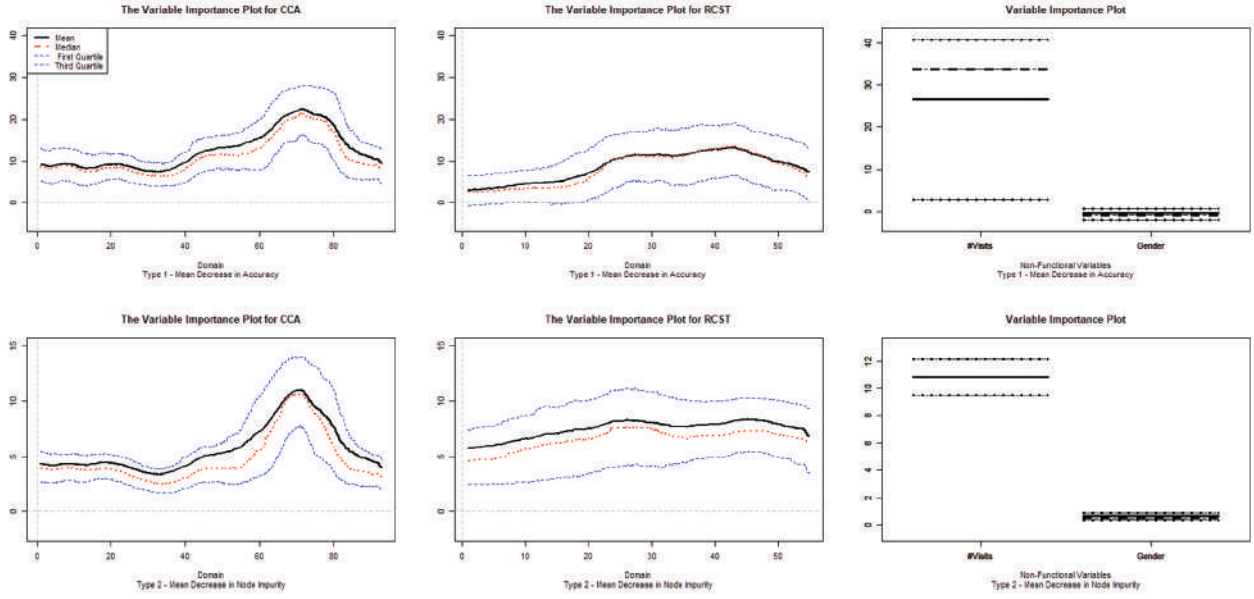

Figure 1. The variable importance plot (VIP) of the MD-RF for CCA, RCST (Functional) and number of Visits and Gender (Non-Functional) group by Type of variable importance.

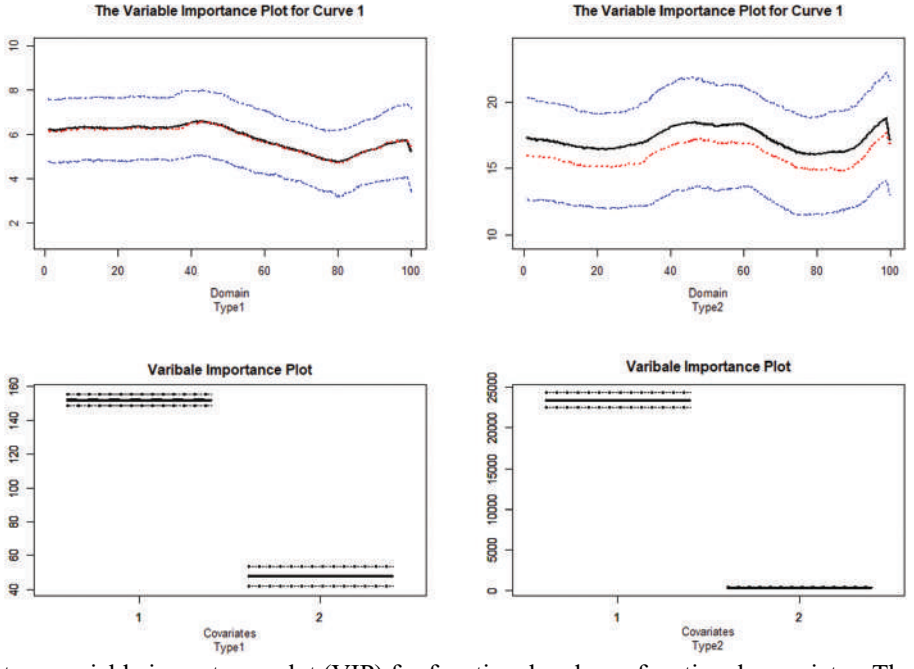

Figure 2_1 : Two type variable importance plot (VIP) for functional and non-functional covariates. The top left is type 1 (mean decrease in MSE) and the top right is type 2 (mean decrease in node impurity) VIP for functional covariate NIT (Curve 1), the bottom left and right are VIPs for non-functional covariates, Water (V1) and Protein (V2) type. Model name is MD-BG. %Fat ~ NIR+Water+Protein. Iterations are 1,000.

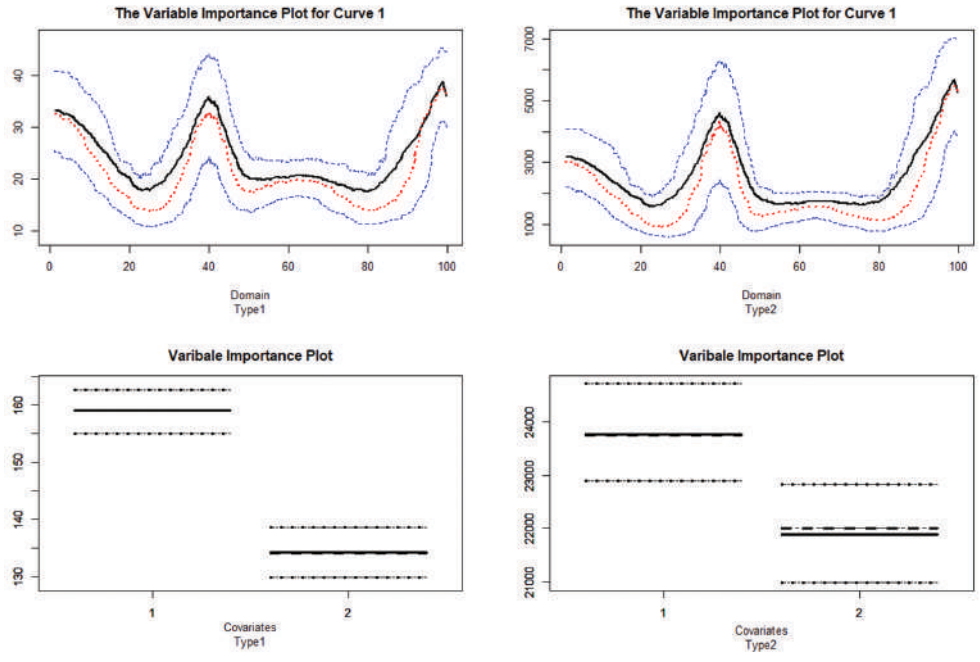

Figure 2_2 : Two type variable importance plot (VIP) for functional and non-functional covariates. The top left is type 1 (mean decrease in MSE) and the top right is type 2 (mean decrease in node impurity) VIP for functional covariate NIT (Curve 1), the bottom left and right are VIPs for non-functional covariates, Water (V1) and Protein (V2) type. Model name is MD-RF. %Fat ~ NIR+Water+Protein. Iterations are 1,000.

Discussion

The proposed method can handle multiple functional covariates each having a different domain along with the nonfunctional continuous and categorical covariates. It also produces a variable importance plot for each covariate and we suggest using MD-RF instead of MD-BG because they do not mask other unimportant variables. We also recommend using MD-BG for the prediction purpose, because the overall model performance is higher than MD-RF. The MD-BG for mixed data has two advantages: 1) it considers each functional covariate as a unit and then converts it to the summary statistics; 2) ACC, SEN, and SPE are the highest among RF, BG, and MD-RF.

The VIP of the MD-BG and MD-RF are smooth and their interpretations are easy as regular BG and RF plus considering the continuous underlying structure of the functional covariates. The main engine of this work is based on the functions from random Forest SRC package, and it adds a random-split procedure for the functional covariate to enhance the model.[18] The R code and examples for reproducing the results are available in the GitHub repository.

The mixed data arise in many statistical models and applications such as precision medicine,[19] electroencephalography (EEG) analysis,[20-22] semi-functional partial linear regression,[23] etc. The mixed data in the BG and RF is considered in this research and one way for the future direction of this research is combining functional principal components analysis (FPCA) and hybrid PCA in the BG and RF models. The R package with vignette is developed in the following address:

mohammad-fayaz/RSRF.

Conclusion

In many applications, there exists many functional and non-functional covariates that have linear and non-linear effects on the response variables. The MD-RF is an extension of RF and it designed for multiple functional and non-functional covariates. It considers each functional covariate as a single entity. The MD-BG is an extension of BG with considering all covariates in the model. This method is developed in the mohammad-fayaz/RSRF, R package.

Conflicts of interests'

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press; 1984.

2. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction: Springer Science & Business Media; 2009.

3. Yu Y, Lambert D. Fitting trees to functional data, with an application to time-of-day patterns. Journal of Computational and graphical Statistics. 1999;8(4):749-62.

4. Febrero Bande M, Oviedo de la Fuente M. Statistical computing in functional data analysis: The R package fda. usc.

5. Gregorutti B, Michel B, Saint-Pierre P. Grouped variable importance with random forests and application to multiple functional data analysis. Computational Statistics & Data Analysis. 2015;90:15-35.

6. Nerini D, Ghattas B. Classifying densities using functional regression trees: Applications in oceanology. Computational Statistics & Data Analysis. 2007;51(10):4984-93.

7. Rahman R, Dhruba SR, Ghosh S, Pal R. Functional random forest with applications in dose-response predictions. Scientific reports. 2019;9(1):1-14.

8. Scornet E. On the asymptotics of random forests. Journal of Multivariate Analysis. 2016;146:72-83.

9. Capitaine L, Bigot J, Thiébaut R, Genuer R. Fréchet random forests for metric space valued regression with non euclidean predictors. 2020.

10. Möller A, Tutz G, Gertheiss J. Random forests for functional covariates. Journal of Chemometrics. 2016;30(12):715-25.

11. Pospisil T, Lee AB. (f) RFCDE: Random Forests for Conditional Density Estimation and Functional Data. arXiv preprint arXiv:190607177. 2019.

12. Silverman B, Ramsay JO. Functional

Data Analysis. 2005.

13. Goldsmith J, Scheipl F, Huang L, Wrobel J, Gellar J, Harezlak J, et al. Refund: Regression with functional data. R package version 01-16. 2016;572.

14. Ferraty F, Vieu P. Nonparametric functional data analysis: theory and practice: Springer Science & Business Media; 2006.

15. Bande MF, de la Fuente MO, Galeano P, Nieto A, Garcia-Portugues E, de la Fuente MMO. Package 'fda. usc'. 2020.

16. RColorBrewer S, Liaw MA. Package 'randomForest'. University of California, Berkeley: Berkeley, CA, USA. 2018.

17. Gareth J, Daniela W, Trevor H, Robert T. An introduction to statistical learning: with applications in R: Spinger; 2013.

18. Ishwaran H, Kogalur UB, Kogalur MUB. Package 'randomForestSRC'. breast. 2022;6:1.

19. Ciarleglio A, Petkova E, Ogden RT, Tarpey T. Treatment decisions based on scalar and functional baseline covariates. Biometrics. 2015;71(4):884-94.

20. Scheffler A, Telesca D, Li Q, Sugar CA, Distefano C, Jeste S, et al. Hybrid principal components analysis for region-referenced longitudinal functional EEG data. Biostatistics. 2020;21(1):139-57.

21. Fayaz M, Abadi A, Khodakarim S. The Functional Regression With Reconstructed Functions From Hybrid Principal Components Analysis: With EEG-fMRI Application. Statistics, Optimization & Information Computing. 2022;10(3):890-903.

22. Fayaz M, Abadi A, Khodakarim S. The Comparison between Visually and Auditory Oddball Tasks in the EEG Experiment with Healthy Subjects. Frontiers in Biomedical Technologies. 2020.

23. Aneiros-Pérez G, Vieu P. Semi-functional partial linear regression. Statistics & Probability Letters. 2006;76(11):1102-10.