DOI: 10.15849/IJASCA.220720.11

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 2, July 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Variable Selection in High Dimensional Data with Interactions

Zuharah Jaafar and Norazlina Ismail
Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Malaysia
zuharahjaafar@gmail.com
Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Malaysia
i-norazlina@utm.my

Abstract

A common research area in statistical machine learning has been variable selection in high dimensional settings. In recent years, numerous effective approaches have been created to deal with these challenges. In order to improve the prediction accuracy of the model for the given dataset, this study sought to present a double approach variable selection method when pairwise interactions between the explanatory variables exist and to choose the smallest explanatory variable set (considering interactions among them). In this study, a double step method consolidating Random Forest and Adaptive Elastic Net was further examined to mimic potential health effects of environmental contamination. When there were existing interactions in the data or none at all, the double step approach was compared to the single-step adaptive elastic net method and two-step CART paired with the adaptive elastic net method. Using significant statistical tests like RMSE, R^2, and the quantity of the variable chosen for the final model, the success of the strategies was measured. The double step RF+AENET approach produces a simple, constrained model. Despite the complex association between exposure variables, it has the lowest false detection rate for null interactions. A set of variables that have correlation with the result are effectively retained by the screening and variable reduction processes in the RF step of the RF+AENET approach. The double step RF+AENET performs prediction better than a single technique and chooses a sparse model that is close to the true model. Thus, it can be said that when there are pairwise interactions between variables in the simulated biological dataset, the double step technique is a better method for model prediction and parameter estimation.

Keywords: Adaptive Elastic Net, Random Forest, Variable Selection, CART.

Received 22 March 2022; Accepted 28 June 2022

1 Introduction

In high-dimensional statistical modelling, variable selection is crucial. It has several modern applications in the fields of medicine, economics, genetics, finance, and many more. A consistent variable selection process and a parsimonious model are the hallmarks of an effective variable selection process. If a number of unimportant variables are chosen, the statistical model may lose its ability to forecast, making it difficult to understand the outcomes.

Additionally, some data reveals strong correlations between predictors. For instance, a typical microarray data typically comprises fewer than 100 samples and thousands of predictors. Due to the challenges of high-dimensional data analysis, existing variable selection approaches primarily concentrate on univariate analysis based on individual influences on the outcome. In order to improve predictions, this study focuses on detecting interactions or synergistic effects between factors. When two or more things or agents work together, the combined effect is higher than the sum of any of their separate additive effects. This is known as a synergist interaction.

The RF technique is used in this study's first stage to minimise the dimension of a number of dependent variables. The link of the covariates with the outcome, plus interaction terms, is quantified in the second phase using an adaptive elastic net. The proposed methods will be investigated further in the analysis of hypothetical simulated health impacts on environmental pollution data, and they will be compared to other methods. The findings showed that, when compared to other methods, the suggested double step approach consistently produces the shortest RMSE and the parsimonious or simplest model. Additionally, it chooses variables that are more akin to the actual model.

2 Related Work

We have recently benefited from several improvements and enhancements in variable selection. The ridge regression developed by Hoerl and Kennard [6] is one of the most popular penalized methods to solve multicollinearity problems. Lasso [13] efficiently chooses variables, estimates significant variable effects, and produces accurate parsimonious models. Lasso is a popular method for simultaneous variable selection and variable effect estimation. However, Lasso cannot select more explanatory variables than the number of observations. Lasso tends to choose one

variable among correlated variables when there is a high correlation among explanatory variables.

Zou and Hastie [19] proposed the elastic net penalty, a combination of the Lasso and Ridge regression penalty, to overcome the disadvantages of the Lasso method. Various applications have proved the advantages of the elastic net, and several properties have been studied [3]. Elastic net performance is better than Lasso in terms of prediction error whenever exists a high correlation between variables [15]. Zou [18] proposed the adaptive Lasso in which adaptive weights are used for penalizing different coefficients in the L1-norm penalty. He concluded the adaptive Lasso could select the model consistently if the weights assigned were a small amount for the essential explanatory variable and a large amount for the unimportant explanatory variables. The adaptive Lasso cannot handle the situation of the existence of highly correlated explanatory variables and unable to select more variables than the number of observations.

As a result, Zou and Zhang [18] proposed the adaptive elastic net by changing the L1-norm penalty with the adaptive Lasso penalty. They proposed to use the elastic net estimates as an initial weight. However, in either $p < n$ or $p > n$, using elastic net estimator as initial weight in adaptive elastic net may not seem suitable. Two reasons are that the Elastic net estimator is inconsistent, which means it is biased in selecting variables. Elastic net can select grouped explanatory variables when the correlation among explanatory variables is more than 0.95. For the first reason, the elastic net does not offer weight for all the explanatory variables, which means that some explanatory variables will be selected, and the others will be set to zero. Based on the second reason, the elastic net cannot encourage the grouping effect if the correlation among explanatory variables does not exceed 0.95.

In data mining and statistical learning, one of the most well-known ensemble learning techniques is random forests (RF). For effective data-adaptive inference, RF, a nonparametric tree-based ensemble approach, combines the concepts of adaptive closest neighbours with bagging. Due to its greedy one-step-at-a-time node splitting and ability to force regularisation for analysis in the p n issue, as well as the grouping property of trees, RF can effectively handle correlation across variables [11]. Additionally, utilising variable importance metrics, RF can be used to choose and rank variables (VIM). As a result, the characteristics of RF make them appropriate for reviewing genomic and bioinformatics studies. Random Forest (RF)

method is one of the machine learning approach that successfully manages datasets with more variables than observations, captures the non-linear relationship between predictor variables, handles missing values, and generates more reliable results that are not affected by the effect of outliers [5].

For the purpose of applying dimension reduction and determining the strength of an association between a variable that includes interaction terms, Li L. [10] presented a double step method consolidating Random forest and ALasso. They concluded that the proposed method outperformed the other techniques by producing a parsimonious model. The proposed method is analyzed on a real Navajo Cohort Birth Study dataset and a simulated dataset.

Sun. Z [11] proposed a two-step approach that combined Classification and regression tree (CART) and variable selection methods to analyze environmental health data. They imposed a two-step strategy by initial screening using a tree-based method followed by five variable selection approaches in the second step. They concluded that there is no method superior to others as the performance differs based on the nature of the response variable, sample size, and the interaction of the variables. Thus, the proposed methods apply to the multi-pollutant framework, and the method used should be based goal of the study, be it a prediction, effect estimation, or screening for significant predictors and their interactions.

Jiali S. [12] proposed a two-stage algorithm based on Least Angle Regression (LARS) and Random Forests. They concluded that the proposed method significantly improved the model fitting and variable selection, requiring less calculation time. The method was applied to analyze real data, which is flowering traits in Arabidopsis and the results showed better performance than others in selection and estimation.

Wang and Zhu [14] developed a two-step approach using the sure independence screening method and adaptive elastic net to perform variable selection in high dimensional regression. They proposed a family of Bayesian information criteria and investigated the selection consistency. Chen et al. [2] combined the ideas of profiling and adaptive elastic net in studying the variable selection in partially linear models. They proved that the proposed method has oracle properties and can handle multicollinearity.

Algamal & Lee [1] proposed a two-stage sparse logistic regression by combining the screening approach as a filter method and adaptive Lasso with a new weight as an embedded method. Their method addresses the effect of high correlation between genes in higher dimensional DNA microarray data. Results revealed that the proposed techniques have a better performance in terms of accuracy in Classification, stability, area under the curve, and G-mean. The result also indicated that the top genes selected are related biologically to the type of cancer. Thus, their method can be applied to the classification of cancer using DNA microarray data in actual clinical practice.

Y. Zhang and Wang [16] examine a combination of two penalized methods, LASSO and Elastic Net, in high dimensional data. An efficient algorithm was employed that will consistently regulate the parameters of LASSO. The two penalized approach has a more significant performance in the prediction and selection of relevant genes in comparison to other methods. Moreover, the results of the study are consistent in several situations. Generally, it has been presented from the study that the LASSO and elastic net model yielded improved predictability of oil price by chosen influential predictors.

3 Methodology

3.1 Penalized regression method

In recent years, an attractive framework of penalized methods has been adapted and gained popularity among statisticians as the key for simultaneously performing variable selection and parameter estimation in high-dimensional data. Consequently, a family of penalized methods was proposed with a penalty term added to the loss function. The advantage behind the penalty term is to control the complexity of the model and provide a criterion for variable selection by introducing some constraints on the coefficients, which force some coefficients to be precisely zero. The amount of the penalty term is the tradeoff between the variance and bias of the selected model [15]. The penalized linear regression, PLR ($\beta; \lambda$) is usually defined as:

$$\text{PLR}\ (\beta; \lambda) = (\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} \lambda\ (\ |\beta_j|\ ), \quad (1)$$

where $p_\lambda(\cdot)$ represents a penalty term, which it is a function of the coefficients, and $\lambda \in [0,\infty]$ is the tuning parameter. The penalty term is fully relied on the λ which controls the amount of the shrinkage. For $\lambda = 0$, we obtained OLS estimates.

However, the larger the values of λ, the influence on shrinkage amount also increased on the coefficient estimates. In penalized linear regression, the coefficient estimates are obtained by minimizing:

$$\hat{\beta}_{\text{PLR}} = \arg min_{\beta} \text{PLR}(\beta; \lambda) \quad (2)$$

3.2 Adaptive elastic net

Elastic net has shown better results than Lasso in many cases with correlated data. On the other hand, the elastic net does not enjoy oracle properties regarding variable selection consistency, although it performs well in prediction accuracy [8]. Zou and Zhang [20] pointed out that the adaptive Lasso outperforms Lasso in terms of enjoying oracle properties even though the high correlation among variables is still a drawback of Lasso. Zou and Zhang [20] proposed an adaptive elastic net (AENET) as an adaptive version of the elastic net to encourage the grouping effect and enjoy the oracle properties simultaneously. Under some assumptions, the adaptive elastic net estimator's consistency was proven. The penalized linear regression using AENET is given by:

$$\text{PLR}(\beta, \lambda_1, \lambda_2)^{AENET} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} w_j |\beta_j| + \lambda_1 \sum_{j=1}^{p} w_j \beta_j \quad (3)$$

The AEN net estimator is then defined as follows:

$$\hat{\beta}^{AEN} = (1+\lambda_2)\,\hat{\beta}_{PLR}^{NAEN} \quad (4)$$

where $\hat{\beta}_{PLR}^{NAEN}$ is the naive adaptive elastic net solution and defined as:

$$\hat{\beta}_{PLR}^{NAEN} = \arg min_{\beta}\{\text{PLR}(\beta; \lambda_1, \lambda_2)^{\text{AEN}}\} \quad (5)$$

3.3 Random forest

The random forest approach is an ensemble-based method that creates a large number of trees. To partition the element space into groups of items with comparative member designs between the indicator factors and the result variable, each tree is assembled using a recurrent distribution approach. Every tree is specifically created by a bootstrap test of dataset that is drawn at random. A model based on the root mean value of the error is used to divide the axes of the trees given a selection of factors that was randomly chosen. By averaging a group of trees, the forecast is created. In the second stage of the study, we preselect significant components using the variable measure of significance (VIMP). VIMP is a ratio of a variable's relevance that assesses the correction of forecast inaccuracy in the event that this variable is left out of the study. A better consistency of a variable is related to a higher VIMP estimate.

3.4 Classification and regression trees (CART)

By partitioning the data space and fitting a straightforward prediction model within each partition, classification & regression trees (CART) is a machine learning approach for creating models from data [9]. The square of the difference of the observation and forecast amount for continuous dependent variables is used to determine the error of prediction on regression trees. The most predictive factors can then be estimated and chosen using this information.

4 Results, Analysis and Discussions

The effectiveness of three different approaches was compared, including the single-step adaptive elastic net (AENET), the double step CART plus the adaptive elastic net method (CART+AENET), and the double step RF plus the adaptive elastic net method (RF+AENET). We created two models for each strategy, one with all of the pairwise interactions between the independent variables and the other with the main effects for the independent variables. 20 correlated exposure variables, X=(X1, X2,...X20), and a continuous outcome variable (Y), which is the result of the linear combination of the main effects and interaction effects of a subset of the exposure variables, make up the simulated dataset that was hypothetically simulated based on real biological data. The sizes 500 and 750 of the samples were also simulated. The simulated data set includes interactions between exposures X1, X2, X12, and X15 that are favourable, interactions between X9 and X16 that are detrimental, and interactions between X1 and X12 that are synergistic. We set the correlation between the variables X1, X2, and X15 to r=0.1 and r=0.05, respectively. A number of performance metrics, including R^2, Adjusted R^2, mean squared error (MSE), and mean squared prediction error (MSPE), were produced to assess how well the techniques performed when analysing biological data (either with or without the interaction terms). To determine the prediction performance measures, we also carried out a 10-fold cross-validation. We specifically separated the datasets into ten parts. Nine partitions served as the training set for the model we built, while the remaining partitions served as the test set for measuring model prediction error. The 10-fold cross-validation prediction error estimates were then calculated by combining the prediction error for each division. The statistical software R was used for all calculations.

4.1 Results of simulated hypothetical biological data (No interaction) for n=500

In Fig. 1 and Table 1, the chosen variables and model performance of the

three techniques are displayed. Using the three methods, Fig. 1's forest chart displays the regression coefficients, link between the exposure factors and the outcome, and their 95% confidence intervals. The single adaptive elastic net approach correctly found null effects for X3 and X7 but correctly identified actual main effects for X1, X2, X9, X12, X15, and X16. Two-step CART+AENET incorrectly recognised two null factors, X3 and X7, and failed to recognise X9 and X15 as real effects. However, the two-step RF+AENET punished the remaining coefficients to zero and selected the majority of the right main effects of the non-zero coefficients. According to Table 1, two-step CART+AENET performs poorly when compared to the other two-step approach, with an adjusted R2 of only 28.2 percent and a greater MSPE of 1.098. The performance of single-step AENET is superior, with an R2 of 30% but a higher MSPE of 1.093. With an R2 of 28.3% but a lower MSPE of 1.075, two-step RF+AENET performs similarly to but better than single-step AENET. The parsimonious model produced by the two-step RF+AENET technique has fewer minor difficulties. Despite the complex association between exposure variables, it has the lowest false detection rate for null interactions. The screening and variable reduction steps in the RF step of the RF+AENET technique successfully retain a set of variables that have any association with the observed values. The two-step RF+AENET strategy performs prediction better than a single approach and has a tendency to choose a parsimonious model that is similar with true model.

Table 1. Performance Evaluation for each methods (No interaction) for n=500

	AENET	**CART+AENET**	**RF+AENET**
R^2	30.0%	28.2%	28.3%
Adjusted R^2	28.9%	27.3%	27.5%
MSE	1.012	1.030	1.039
MSE.CV	1.007	1.038	1.036
MSPE.CV	1.093	1.098	1.075

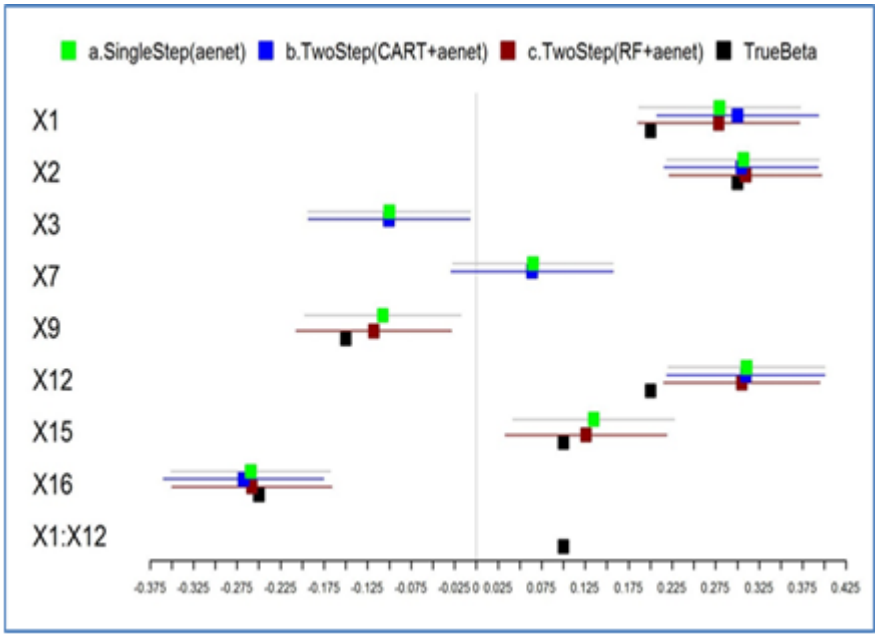

Fig 1. Analysis of Simulated Dataset (with interaction) for n=500.

The forest plot describes the regression coefficients (box) and 95% confidence interval (line) on modeling the relationship between exposure variables and the outcome using methods: AENET (green), 2) CART+AENET (blue), 3) RF+AENET (dark red). The black box indicate the true parameter of coefficient.

In a subsequent simulation, we employ increasing correlation (r=0.3, 0.5, and 0.7) to further assess the model's effectiveness utilising various correlation strengths among the exposure variables. On the various correlation levels, the exact amount of variables (m=20) and n=500 were simulated. In each round of rising correlations, the estimated regression coefficients were compared to the actual simulation coefficients. Table 2 compares the results of the performance evaluations of each correlation setting using the three different approaches. The results from Table 2 demonstrate that the RF+AENET performs best in the areas where its MSPE is the least. Additionally, it has superior prediction power and chooses fewer variables that are closer to the actual model. The RF+AENET solution that was suggested consistently performed better than the other two methods. Both the two-step CART+ AENET and the single-step adaptive elastic net choose more variables than the actual model does. Regarding the strength of the correlation, the three methods consistently produce better results as the correlation increases (r>0.7), but they did poorly at correlations higher than that, where they mistakenly thought that the variables were null and included them in the final model.

Table 2. Performance Evaluation for each method on different degree of correlations for n=500

	AENET	CART+AENET	RF+AENET
r = 0.3			
R^2	36.3%	33.7%	32.1%
MSPE.CV	1.280	1.146	1.115
r = 0.5			
R^2	37.3%	38.6%	34.5%
MSPE.CV	1.166	1.109	1.103
r = 0.7			
R^2	49.7%	43.4%	46.6%
MSPE.CV	1.147	1.164	1.113

4.2 Results of simulated hypothetical biological data (With interaction) for n=500

Figure 2 and Table 3 display the chosen variables and model performance of the three approaches, which simulate the main effects and pairwise interactions of the exposure variables. Based on Fig. 2, the two-step RF + AENET identified the impacts of the X1 and X12 interactions and produced the parsimonious model. The complex correlation structure between the variables may have had an impact on the selection of a few null interaction variables by all three models. However, RF+ AENET is thought to have the lowest false detection rate among the three approaches, which makes it the best. The model produced by the single-step AENET had the highest adjusted R2 and contained more interaction effects than other models. Overfitting, which via cross-validation contributed to the greatest MSPE, is most likely to blame for this. When compared to other approaches, CART+AENET has the greatest MSE and the smallest adjusted R2. With the smallest MSPE, RF+AENET produced better results and chose the sparsest, most minimal model that was closest to the real model.

Table 3. Performance Evaluation for each methods (with interaction) for n=500

	AENET	CART+AENET	RF+AENET
R^2	34.2%	31.6%	31.8%
Adjusted R2	31.2%	29.5%	29.7%
MSE	0.952	0.990	0.987
MSE.CV	0.901	0.995	0.982
MSPE.CV	1.234	1.106	1.104

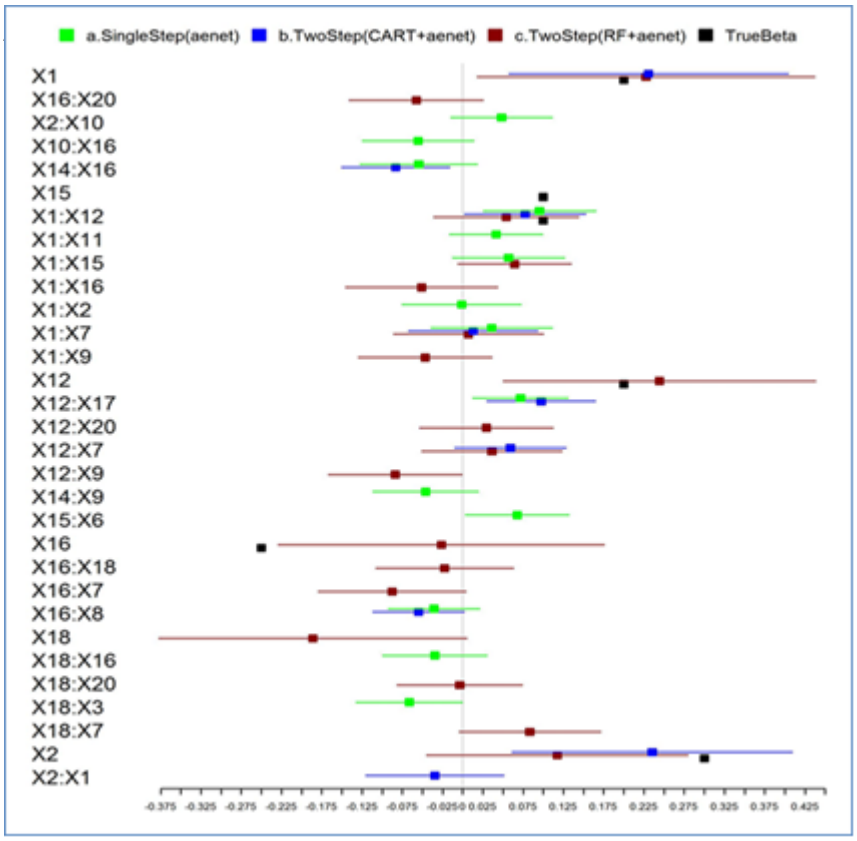

Fig. 2. Analysis of Simulated Dataset (with interaction) for n=500.

4.3 Results of simulated hypothetical biological data (No interaction) for n=750

Table 4 displays the model performance for the three methods. The single adaptive elastic net approach incorrectly detected null effects for X3, X4, X7, X11, and X15 while correctly identifying actual main effects in X1, X2, X9, X12, X15, and X16. Two-step CART+AENET incorrectly selected two null variables, X7 and X11, as real effects in place of X15. However, the two-step RF+AENET punished the remaining coefficients to zero and selected the majority of the right main effects of the non-zero coefficients. In contrast to the other two-step technique, two-step CART+AENET performs poorly based on Table 4's smaller adjusted R2 of 29.4% and bigger MSPE of 0.907. The performance of single-step AENET is superior, with an R2 of 30.5 percent and an MSPE of 0.809. With an R2 of 29.9% and the smallest MSPE of 0.896, two-step RF+AENET performs similarly to but better than single-step AENET. The two-step RF+AENET approach produces a simple, constrained model. Despite the complex association between the exposure variables, it has the lowest false detection rate for null interactions. The screening and variable reduction steps in the RF step of the RF+AENET technique successfully retain a set of variables that have any association with the outcome.

Table 4. Performance Evaluation for each methods (No interaction) for n=750

	AENET	CART+ AENET	RF+AENET
R^2	37.2%	31.7%	32.5%
Adjusted R^2	32.7%	29.4%	30.6%
MSE	0.790	0.858	0.849
MSE.CV	0.779	0.864	0.848
MSPE.CV	1.043	0.938	0.928

4.4 Results of simulated hypothetical biological data (With interaction) for n=750

The parsimonious model was produced via the two-step RF + AENET, which identified the interaction effects between X1 and X12. The complex correlation structure between the variables may have had an impact on the selection of a few null interaction variables by all three models. However, RF+ AENET is thought to have the lowest false detection rate among the three approaches, which makes it the best. The model produced by the single-step AENET had the highest adjusted R2 and contained more interaction effects than other models. Overfitting, which via cross-validation contributed to the greatest MSPE, is most likely to blame for this. In comparison to other approaches, CART+AENET yields the smallest adjusted R2 and the average MSE. With the smallest MSPE, RF+AENET produced better results and chose the sparsest, most minimal model that was closest to the real model.

Table 5. Performance Evaluation for each methods (with interaction) for n=750

	AENET	CART+ AENET	RF+AENET
R^2	37.2%	31.7%	32.5%
Adjusted R^2	32.7%	29.4%	30.6%
MSE	0.790	0.858	0.849
MSE.CV	0.779	0.864	0.848
MSPE.CV	1.043	0.938	0.928

4.5 Discussion of results

In order to investigate the link between exposure variables and the outcome in this study, including pairwise interaction effects and main effects, we created a two-step methodology integrating the Random Forest (RF) and adaptive elastic net (AENET). The goal of this work was to enhance the two-step CART+ LASSO methodology that was previously put forth [11]. A single tree technique with some shortcomings is the CART. First off, the tree is frequently non-robust, meaning that even a small modification to the training dataset might have a significant impact on the trees' predictions. Second, while the CART approach does not ensure perfect global trees, it produces ideal local trees. Last but not least, overfitting caused by CART could result in extremely complex trees that do not generalise well to other test data.

Due to the randomness included in the algorithm and the use of "out of bag"

predictions to evaluate the model performance, the RF approaches suggested in the first stage are employed for model dimension reduction and can produce more reliable results than CART. Then, we choose and penalise additional variables using the adaptive elastic net (AENET) approach. A single-step adaptive elastic net was shown to choose more variables than the actual model. As a result, the data dimensionality can be decreased at the initial step of employing RF. The RF technique involves choosing covariates that are related to the outcome, whether there is a linear or non-linear relationship between them, and taking into account their interactions.

The adaptive elastic net approach seeks to find a sparse model with higher prediction performance after the pre-filtering variables step. The double step strategy addressed the shortcomings of the CART and Lasso approach, which always yields biassed estimates for big coefficients or when there is multicollinearity within the variables. It also contributed a practical weighted penalty to the coefficients. The adaptive elastic net method addresses LASSO's flaws, which include its propensity to select a group of highly correlated variables for the model by encouraging grouping effects.

4.6 Software

Rstudio 4.0.5 was used to conduct all statistical computations and simulations. The "glmnet" package in R was used to fit AENET models, while the "randomForestSRC" library was used to implement RF model selection. In this study, we focus on evaluating the efficacy of the suggested strategy using simulated datasets that mimicked real-life biological data where there is pairwise interaction between covariates.

5 Conclusions

All three models frequently selected a small number of null interaction variables in tests, which may have been influenced by the intricate correlation structure between the variables. However, RF+ AENET is thought to have the lowest false detection rate among the three approaches, which makes it the best. The model produced by the single-step AENET had the highest adjusted R2 and contained more interaction effects than other models. Overfitting, which via cross-validation contributed to the greatest MSPE, is most likely to blame for this. When compared to other approaches, CART+AENET has the greatest MSE and the smallest adjusted R2. With the smallest MSPE, RF+AENET produced better results and chose the sparsest, most minimal model that was closest to the real model. As a result, it can be said that the two-step strategy combining Random Forest (RF) and adaptive elastic net (AENET) outperforms both the single-step AENET and the two-step CART + AENET in terms of performance. Using a simulation dataset, this approach consistently identifies complicated interactions between exposures. The method also produces a sparse model that is comparable to the actual model and performs predictions more accurately than the single-step method. It is advised that in the future, in order to cope with strongly correlated variables in high-dimensional settings, another method that has been demonstrated to be effective in one-step estimation, such as Mnet penalty, weigh-fused elastic net penalty, and others, be investigated.

ACKNOWLEDGEMENTS

The authors would like to thank the editor and anonymous reviewers for their insightful comments and suggestions. It helped to improve the quality and composition of this paper. And also special thank you for R package "glmnet" and "randomForestSRC" within the R.

REFERENCES

[1] Algamal, Z. Y., & Lee, M. H. (2019). A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Advances in Data Analysis and Classification, 13(3), 753–771.

[2] Chen, B., Yu, Y.,Zou, H. and Liang, H. (2012). Profiled adaptive Elastic Net procedure for partially linear models with high-dimensional covariates. Journal of Statistical Planning and Inference. 142(7), 1733-1745.

[3] De Mol, C., De Vito, E. and Rosasco, L. (2009) Elastic-net Regularization in learning theory. Journal of Complexity. 25(2), 201-230.

[4] Fan, J., & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348–1360.

[5] Guo L, Ma Y, Cukic B, Singh H. (2004) Robust prediction of falt-proneness by random forests. 15th international symposium on software reliability engineering. Proceeding. 417-28.

[6] Hoerl, A. E., & Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics,12,55–67.

[7] James, G. Witten, D., Hastie, T. and Tibshirani, R. (2013) An introduction to statistical learning with applications in R. NY:Springer.

[8] Leng, C., Lin, Y. and Wahba, G. (2006) Note on the lasso and related procedures in model selection. Statistica Sinica. 21, 391-419.

[9] Loh WY. (2011) Classification and regression trees. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery.1(1):14-23.

[10] Luo L, Hudson LG, Lewis J, Lee JH. (2019) Two-step approach for assessing the health effects of environmental chemical mixtures: application to simulated datasets and real data from the Navajo Birth Cohort Study. Environ Health.

[11] Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, Batterman SA, Mukherjee B. (2013) Statistical strategies for constructing health risk models with multiple pollutants and their interaction: possible choices and comparisons."Environ. Health. 12(1):85.

[12] Sun, J., Wu, Q., Shen, D. et al. (2019) TSLRF: Two-Stage Algorithm Based on Least Angle Regression and Random Forest in genome-wide association studies. Sci Rep 9, 18034.

[13] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal

of the Royal Statistical Society. Series B (Methodological), 5, 267–288.

[14] Wang, T. and Zhu, L. (2011) Consistent tuning parameter selection in high dimensional sparse linear regression. Journal of Multivariate Analysis. 102(7), 1141-1151.

[15] Yang, H., Guo, C. and Lv, J. (2015) SCAD penalized rank regression with a diverging number of parameters. Journal of Multivariate Analysis. 133, 321-333.

[16] Zhang, Y., Ma, F., & Wang, Y. (2019). Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? Journal of Empirical Finance, 54(September), 97–117.

[17] Zhou, D.-X.(2013) On grouping effect of elastic net. Statistic & Probability Letters. 83(9), 2108-2112.

[18] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, 101, 1418–1429.

[19] Zou, H., & Hastie, T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 301–320.

[20] Zou, H. and Zhang, H (2009). On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics. 37(4), 1733-1751.

Notes on contributors

Zuharah Jaafar received her Bachelor and Master Degrees in Mathematics from Universiti Teknologi Malaysia in 2007 and 2011, respectively. Currently, she is pursuing PHD in Mathematics at Universiti Teknologi Malaysia. Her main research interest is Machine Learning and she has strong background in Statistics and mathematical computing. She has published two conference papers on variable selection and machine learning methods.

Norazlina Ismail received her Bachelor and Master Degrees in Mathematics from Universiti Kebangsaan Malaysia in 2000 and 2002, respectively. She received her Phd in 2014 from Massey University, New Zealand. From 2000 to 2002, she worked as a tutor, in the Faculty of Science, Universiti Teknologi Malaysia. Currently, she works as lecturer in the Department of Mathematics, Faculty of Science, University Teknologi Malaysia.