International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Special Issue 28 (August, 2015), PP.48-53

A HYBRID RANDOM FORESTS-BORUTA FEATURE SELECTION ALGORITHM FOR BIODEGRADIBILITY PREDICTION

Zhe F. Liu[1], Hedia Fgaier[2], Stanislav Y. Ivanov[3¶], Ali Elkamel[1], Xiang H. Meng[4], and Suo Q. Zhao[4]
[1]University of Waterloo, Waterloo, Canada
[2]University of Guelph, Guelph, Canada
[3,2¶]University of Western Ontario, London, Canada
[4]China University of Petroleum (Beijing), Beijing, China

***Abstract*— The a priori knowledge about biodegradability is adopted to save time and money for research and design of new products. Quantitative structure activity relationship (QSAR) models as a tool for biodegradability prediction of chemicals have been encouraged by environmental organizations. In the current work, a new algorithm has been proposed to investigate the importance of chemical descriptors to be used as input variables in modeling and prediction of biodegradability. The algorithm allows obtaining an ensemble of feature subsets compromising between model complexity and generalization performance. It utilizes random forests as classifier coupled with Boruta algorithm to automatically rank and omit descriptors based on Z-score. It is shown how four least relevant variables were identified and removed from model remaining generation ability. Furthermore, a hybrid feature selection method is developed to inspect weak relevant features and omit them in a loop mode in order to remain generalization of classifiers. The prediction accuracy of the new model showed improvements compared to previous works.**

Index terms—**QSAR; Random forests; Boruta, Hybrid feature selectio, optimization**

I. INTRODUCTION

Modern society puts a lot of effort into keeping its environment safe and clean. At the same time a wide variety of present-day consumer products are composed of different chemicals which might pose a threat to the nature due to their accumulation and persistence in the soil, water or air. The number of goods and their turnover is drastically increasing thus making the issue of chemicals disposal even more vital. One of the possible ways to reduce the impact on the environment is to use biodegradable chemicals in design and production of new materials[1]. On one hand, one can use experimental methods to define whether the designated chemical is biodegradable or not but this becomes more and more impossible task since tens of thousands of compounds have to be tested. The other group of approaches called quantitative structure activity relationship (QSAR) and quantitative structure property relationship (QSPR) models have been utilized to predict biodegradability[2]. They are based on classification or regression methods, respectively, exploiting properties of molecules. They often utilize a high number of descriptors (variables) for prediction[3]. The variables in highly dimensional model input are often intercorrelated or are not all relevant to the dependent variables, which would deteriorate the performance of QSAR/QSPR models[4].

In this light, feature selection is an important issue to be solved in order to achieve proper model prediction. It can delete redundant descriptors to improve computing speed, save storage, and also enhance transparency of data structure via filter, wrapper and embedded solutions[5]. Filter techniques obtain the relevance of variables and remove unimportant ones from model inputs. Although they lack consideration about relations between features and are non-specific to prediction method, filter-based group of methods reduce model complexity while preserving satisfactory generalization ability; moreover, they are highly computationally efficient[6]. Wrappers allow avoiding problems of filter methods at some extent but with the cost of high computational time. Embedded method can build an optimal subset of features search process into classifiers construction. It is also a time costing method but less than wrapper.

Mansouri et al.[7] implemented a wrapper combining genetic algorithm with a number of machine learning methods (kNN, PLSDA, SVM). The results have showed reduction of the number of descriptors from 781 to 12, 23 and 14 respectively providing satisfactory accuracy of prediction. Although SVM and GA coupled demonstrated the best performance, learning of SVM is parameter sensitive thus its optimum search is harder. Moreover, training of single SVM is a lengthy process itself, so when combined with GA for feature selection it becomes the most time consuming method. Rudnicki et at.[8] adopted Boruta algorithm[9] to search all-relevant features including 37 descriptors of chemical biodegradability data. One of the major drawbacks of this approach is that Boruta method cannot always completely separate variables into relevant or irrelevant leaving some of them without any decision (so called "tentative" variables). Cao and Leung[10] incorporated a wrapped support vector classifier into a differential

evolution algorithm (DE-SVC) and tested its performance on the data set of Mansouri et al. They reported improved performance of classifier compared to the original work.

In this work a new method for feature selection is proposed. It combines filtering and wrapping features thereby attenuating drawbacks of these methods independently. The method is founded on principles of Random Forest (RF)[11] and Boruta algorithm and is aimed for search of relevant subset of features without losing generalization ability. Random Forest (RF) is used as classifier and Boruta model based on the multiple runs of RF is adopted as a filter built in a loop to delete unimportant features. The details are given in the Modeling Methods section below.

- MODELING METHODS
  - *Random Forests*

Random Forests[11] is an ensemble of tree predictors for classification and regression. A subset of features is selected randomly to be used as descriptors included in each tree in RF in order to ensure the diversity in ensemble of trees, which is the key to why generation error improvements could be obtained by ensemble classifiers. Then a prediction of each tree in ensemble is taken into account by averaging them which yields to the final prediction.

$$P(\overline{x}) = \frac{\sum_{i=1}^{N} P_i(\overline{x_i})}{N} \quad (1)$$

where P is final prediction, $\overline{x}$ is input vector to RF, P_i is a prediction of i^{th} tree in RF, $\overline{x}_i$ – input vector i^{th} into tree of RF consisting of randomly selected subset of descriptors, N – number of trees in RF. If number of trees N is high enough then averaging a solution can provide a satisfactory convergence of RF to predicted value.

Each tree in RF is built using different training sets selected by bootstrapping technique. About one third of whole data[12] are never trained by "tree" classifiers, which constitute an Out Of Bag (OOB) set. OOB accuracy can be used to assess the performance of predictors and variable importance (VI) built in RF algorithm. This results in another advantage of RF as a bagged predictor, which gives it the capability to deal with overfitting.

There are two ways to evaluate variable importance (VI) built in RF. First is to compute from permutations of OOB data as shown below. Initially, OOB prediction error of each tree is recorded, then one feature (say i^{th} molecule descriptor) from OOB data is randomly permuted and new error is recorded. The difference of these computations of OOB errors is averaged between all trees. In such a manner dividing the standard deviation of the difference for all trees one can give the importance value of permuted feature known as Z-score. This way of VI measurement is adopted in this paper since it allows avoiding lengthy cross-validation. Another VI measure – Gini index – is not considered in this work.

The variable importance (VI) for tree t is given by:

$$VI^{(t)}(X_j) = \frac{\sum_{i \in \overline{OOB}^{(t)}} I(y_i = \widehat{y}_i^{(t)})}{|\, OOB^{(t)}|} - \frac{\sum_{t \in \overline{OOB}^{(t)}} I(y_i = \widehat{y}_{i,\Pi_j}^{(t)})}{|\, OOB^{(t)}|} \quad (2)$$

where $\widehat{y}_i^{(t)}$ and $\widehat{y}_{i,\Pi_j}^{(t)}$ represent, respectively, class prediction before and after permutation of variable X_j, $\overline{OOB}^{(t)}$ is the out-of-bag sample for a tree t. Then it is averaged for all trees:

$$VI(X_j) = \frac{\sum_{t=1}^{Number\ of\ trees} VI^{(t)}(X_j)}{Number\ of\ trees} \quad (3)$$

Finally Z-score of variable j is found as:

$$Z_j = \frac{VI(X_j)}{\sigma / \sqrt{Number\ of\ trees}} \quad (4)$$

where σ is standard deviation of $VI^{(t)}(X_j)$.

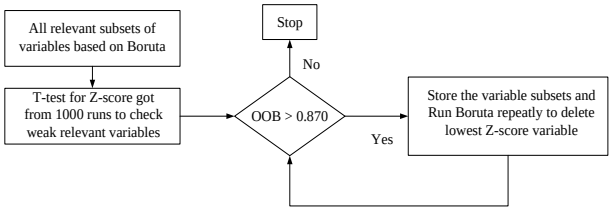
  - *Boruta Algorithm*

Boruta algorithm works in a way that it randomly generates copies of attributes for all inputs (called shadow attributes) and compares input importance value obtained from RF between real and shadow ones. The suggested model only keeps inputs that overwhelm shadow attributes[9]. In order to avoid instability for a single RF run, Boruta adopts multiple computations and compares Z-score values distribution for all variables. Only the values of original features which are higher significantly in statistics than the highest Z-score of permuted attribute (HZPA) are considered as important ones while others are suggested to be removed from the model.

  - *Hybrid Feature Selection*

The proposed new method (Figure 1) includes three steps to obtain a simple subset of descriptors without losing too much prediction ability for biodegradability. A novel filter feature selection method in a loop mode using median value of Z-score as criteria of variable importance is developed in the third step after omitting non-relevant and weak relevant features.

Flow Chart of Proposed Hybrid Feature Selection Algorithm



On the first step of feature selection, Boruta algorithm is used to eliminate non-relevant variables among 41 descriptors. At the next step, the number of RF runs is increased to 1000 aiming to verify whether Z-score of inputs are significantly higher than of shadow ones. Thirdly, a novel filter is developed to omit unimportant descriptors one by one in each run. The process is repeated until the OOB accuracy gets below the threshold value. Here the value of 0.870 is chosen as threshold since it is close to 0.875 that is the value of OOB prediction accuracy using the

original descriptor set. The reason of using Z-score to represent variable importance is to consider variance of feature importance gained from all trees in RF. Nevertheless, there are still chances for weak features to be ranked in an advanced place only based on single run outcome. Hence, median value of Z-score obtained from multiple runs of RF can decrease instability of feature importance. Meanwhile, Boruta algorithm incorporates multiple runs of RF internally, which can be coupled with RF to constitute a feature selection without excessive coding. To summarize, the combination of Random Forest and Boruta algorithm as a filter method can avoid long-run computations issues in wrapper while remaining stable due to implementation of Z-score attained from multiple RF runs in Boruta. In addition, an ensemble of descriptors' subsets can be obtained in the end of computation loop, which provides more opportunities for deciders to comprise between model complexity and its generation ability. Flow chart of entire research is displayed in figure 1.

- METERIALS AND SOFTWARE

The data set of Mansouri et al.[7] is utilized for modeling purposes in this work. The data set consist of 1725 chemical molecules and 41 descriptors are used to represent each molecule. Training, testing and external validation sets have been selected randomly to include 837, 218 and 670 molecules, respectively. The numbers of ready biodegradable (RB) and not ready biodegradable (NRB) molecules in these data sets are shown in Table I while symbols and brief demonstration of molecular descriptors can be found elsewhere (Mansouri et al., 2013).

Numbers of Molecules Included in Different Data Sets

Data	Ready biodegradable	Not ready biodegradable
Training set	284	553
Testing set	72	146
External validation set	191	479

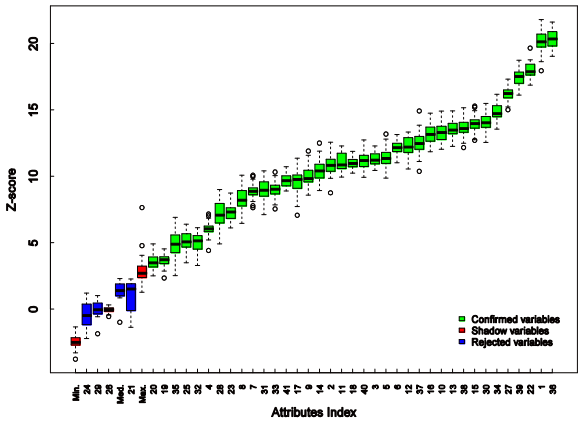
The research was implemented in the R programming language[13]. The packages used in R are Boruta[14]for Boruta algorithm, randomForest[15] for Random Forests, doMC[16] and foreach[17] for multi-core parallel computing. Ubuntu 14.04 LTS was adopted as computer operation system.

- Results and discussion

Boruta algorithm was utilized to test the importance of 41 descriptors in 837 molecules. The parameters of algorithm with respect to RF were set to defaults which are 500 for number of trees and 6 for number of variables to sample at each split. A P-value of 0.01 for Boruta was used to determine the confidence interval of variable importance. The box-plot of variable importance represented by Z-score value is shown in Figure 2.

The entire computation process comprising of 30 iterations was completed in 1.6 minutes. Four descriptors including B01.C.Br, B04.C.Br, N.073 and nCRX3 displayed as blue boxes in Figure 2 are classified as noise variables by Boruta since their Z-score is significantly lower than HZPA shown in the most right red box. The remaining 37 features combine all relevant variable for biodegradability (green box).

Box Plot of Variable Importance Including Shadow Feature



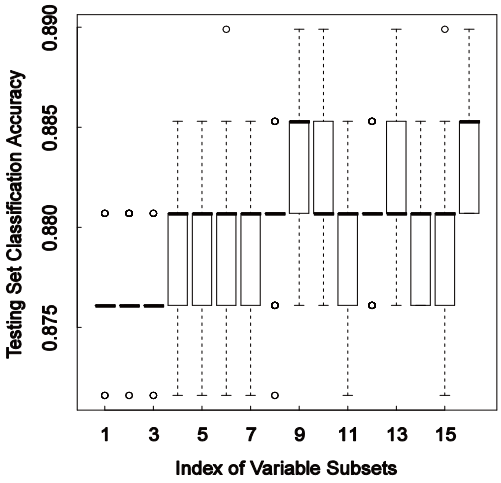
- *T-test for All Relevant Variables*

In order to inspect and remove weakly relevant variables among the remaining 37 descriptors, 1000 runs of Z-score were computed and compared with the highest Z-score shadow feature (HZSF). Two most unimportant descriptors (nN.N and nArNO2) were taken to perform T-test with HZSF respectively yielding results of 30.59 and 24.76 respectively for each of descriptors. These values are much larger than the value of 3, which is the threshold value proposed by Rudnicki et al.[18]. Hence we are able to conclude that even though the most unimportant variables are much more significant than HZSF, all the variables are considered as important.

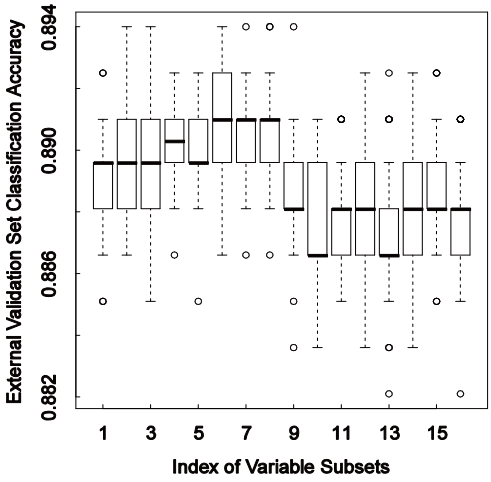
- *Hybrid Feature Selection Coupled Random Forests with Boruta Algorithm*

In this section a filter section coupled with Boruta algorithm as described above was used to search for satisfactory variable subsets at a more detailed level. In a loop process one candidate of variable subset at a time is introduced into Boruta algorithm, where it undergoes multiple runs of RF. Then Z-scores for all variables are taken. Based on median value of Z-score one can rank features and remove the least important descriptor. After it is eliminated, a new subset of variables (same as original but short on the one with the lowest Z-score) is imported into RF model to calculate OOB classification accuracy for 50 times. When if OOB accuracy is higher than predefined threshold value, a current subset of features is stored and sent again into Boruta model to remove the next least important variable and check with OOB accuracy for a new subset. The loop repeats until OOB prediction drops below the threshold value. The approach yields a subset of relevant variables which can be used for model validation on testing and external validation sets.

Variable Subsets in Ensemble Performance on Testing Data Sets



Variable Subsets in Ensemble Performance on External Validation Data Sets



Finally, the decision maker can compromise between prediction accuracy and complexity of model to select a satisfactory subset of features. Details of feature subsets are described in Table II. The first subset (Fig. 3, 4) was obtained from Boruta model after removal of four overshadowed variables from the original 41 variables. The third column of Table II shows the order of variables removal based on Z-score value. At each run of filter loop, one descriptor with lowest median Z-score is omitted until OOB accuracy is lower than threshold. The smallest subset of descriptors obtained by this procedure contains 22 variables which is almost a half of original number of descriptors.

Part of Ensemble Feature Subsets Description

Index of feature subsets	Subsets size	Deleted index of variables from 41 original descriptors
1	37	24, 29, 26, 21
2	36	24, 29, 26, 21, 20
5	33	24, 29, 26, 21, 20, 19, 35, 25
8	30	24, 29, 26, 21, 20, 19, 35, 25, 32, 4, 23
11	27	24, 29, 26, 21, 20, 19, 35, 25, 32, 4, 23, 28, 8, 7
14	24	24, 29, 26, 21, 20, 19, 35, 25, 32, 4, 23, 28, 8, 7, 31, 33, 41
16	22	24, 29, 26, 21, 20, 19, 35, 25, 32, 4, 23, 28, 8, 7, 31, 33, 41, 17, 9

Figures 3 and 4 show that the eighth subset of variables (having 30 variables) performs better for both testing and external validation sets than the first subset (with 37 variables) which has similar generalization ability with original descriptor set (Table III). The generalization performance among all of the subset candidates does not vary significantly, thus the decision maker can also choose the simplest subset as input vector for RB/NRB classification. The performance of the original set, first, eighth and last candidates of ensemble are compared with the results obtained by former researchers and are summarized in Table III.

In Table III three evaluation criteria have been used, which are classification accuracy (Acc), sensitivity (Sn) and specificity (Sp). They are defined as follow:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$Sn = \frac{TP}{TP + FN} \quad (6)$$

$$Sp = \frac{TN}{TN + FP} \quad (7)$$

where, TN and FN are the values of true negative and false negative, and TP and FP are the values of true positive and false positive.

Performance of Part of Candidates in Ensemble Compared with Results of Former Research on Testing and External VALIDATION Sets

Models	Testing set			External Validation set		
	Acc	Sn	Sp	Acc	Sn	Sp
kNN [a]	0.85	0.81	0.90	0.83	0.75	0.91
PLSDA [a]	0.85	0.83	0.87	0.83	0.80	0.86
SVM [a]	0.86	0.82	0.91	0.82	0.74	0.91
DE-SVC [b]	0.877*	0.77	0.93	0.877**	0.74	0.93
RF [c]	0.876	0.736	0.945	0.890	0.754	0.946
1st subset	0.876	0.736	0.945	0.889	0.749	0.945
8th subset	0.881	0.736	0.952	0.891	0.754	0.946
16th subset	0.885	0.750	0.945	0.888	0.738	0.946

a results of Mansouri et al., b is from Cao. et at., c is from random forests with 41 descriptors.
* from Cao. et at. (2014) Table 3. ** from Cao. et at. (2014) Figure 3

It can be observed from the table that the last five classifiers (4 of them use random forest, 1 uses DE-SVC) outperform the first 3 classifiers described by Mansouri et al. Comparing results of 1st subset in ensemble without four overshadowed variables deleted by Boruta with the one of RF with 41 features, one can recognize that there is a slight difference between the two classifiers, which indicate that unimportant variables identified by Boruta can be omitted safely. This fact means that the robustness of random forest method is satisfied as described by former research [19]. The 8th subset with 30 descriptors performs best on external validation set with a total accuracy of 0.891, which is the most important index to estimate classifiers since the data set including 670 data points is much larger than the testing set. After removing 11 descriptors the generalization ability even increases than before, which is a proof of existence of weak relevant features in this data structure and this fact can deteriorate the performance of classifiers. Even 16th subset with just 22 variables is also a competitive candidate for building of classifier for biodegradability prediction as it shows similar performance. Among all of the classifiers considered, the best performance is shown by the 8th subset selected by the new algorithm. Overall it has yielded in 1.4 % (0.891 vs. 0.877 prediction accuracy) improvement compared to the best known classifier. Considering the huge amount of molecules considered, this is a satisfactory outcome. As for the average runtime, the proposed algorithm for the purpose of this paper is 576.4s using 5 clusters of CPU simultaneously. The utilized CPU is Intel® core (TM) i7-3632 QM 2.2 GHz. Since the running time of SVM coupled with Genetic Algorithm (GA) wrapper feature selection cannot be found in Mansouri et.al paper. We run this method on the same computer using 5 as cost parameter in SVM based on package e1071[20] and GA[21]. The parameters' values of GA were used as 50 for population, 200 for generations. Crossover and mutation possibility are 0.8 and 0.2 respectively. The final feature subset is the same as that of the work of Mansouri et.al but whole feature selection process costs 3728.3s, which is more than 4 times longer than the algorithm of this paper.

II. Conclusion

The purpose of this work was to propose an improved version of feature selection technique in order to find an ensemble of features which can improve prediction of RB/NRB taking into account the simplicity of models. Initially algorithm selects unimportant features based on Boruta algorithm. On this step 4 variables out of 41 were rejected as unimportant. Then it verifies the importance of remaining variables by performing T-test on multiple run of Boruta test. The results revealed that even the two least important descriptors belong to all relevant features in the data structure. Finally, an ensemble of 16 variable subsets was obtained by calculation of an OOB score of prediction with the remaining variables. One of them with 30 descriptors shows best generation ability reaching a classification accuracy of 0.891 on external testing sets that contain 670 molecule samples. While a subset with least number of variables still possess satisfactory accuracy and it is qualified to be considered by decision maker. We can summarize that the use of random forest in classification for biodegradability prediction shows comparable results with Cao and Leung[10]. Combining the proposed novel algorithm of feature selection and random forest classifier we were able to achieve even better results by 1.4 % comparing to the best results reported up to this date.

Acknowledgment
The first author would like to acknowledge support from China Scholarship Council. The authors would like to also acknowledge partial support from the Natural Science and Engineering Research Council of Canada.

References

R. S. Boethling, "Designing Biodegradable Chemicals," in Designing Safer Chemicals, vol. 640, American Chemical Society, 1996, pp. 156–171 SE – 8.
B. Philipp, M. Hoff, F. Germa, B. Schink, D. Beimborn, and V. Mersch-Sundermann, "Biochemical interpretation of quantitative structure-activity relationships (QSAR) for biodegradation of N-heterocycles: A complementary approach to predict biodegradability," Environ. Sci. Technol., vol. 41, no. 4, pp. 1390–1398, 2007.
M. Shahlaei, "Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study," Chem. Rev., vol. 113, no. 10, pp. 8093–8103, 2013.
F. Emmert-Streib, M. Dehmer, K. Varmuza, and D. Bonchev, Statistical modelling of molecular descriptors in QSAR/QSPR. John Wiley & Sons, 2012.

I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.

R. Kohavi and G. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 97, pp. 273–324, 1997.

K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, "Quantitative structure-activity relationship models for ready biodegradability of chemicals," J. Chem. Inf. Model., vol. 53, no. 4, pp. 867–878, 2013.

P. W. Rudnicki W.R., Wrzesień M., Feature selection for data and pattern classification. Springer, pp. 11-28, 2013.

M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta - A system for feature selection," Fundam. Informaticae, vol. 101, pp. 271–285, 2010.

Q. Cao and K. M. Leung, "Prediction of Chemical Biodegradability Using Support Vector Classifier Optimized with Differential Evolution," J. Chem. Inf. Model., vol. 54, no. 9, pp. 2515-2523, 2014.

L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

L. Breiman, "Out-of-bag estimation," Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708, 1996b. 33, 34., 1996.

R Development Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.," R Found. Stat. Comput. Vienna, Austria., 2012.

M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," J. Stat. Softw., vol. 36, no. 11, pp. 1–13, 2010.

A. Liaw and M. Wiener, "Classification and Regression by randomForest," R news, vol. 2, pp. 18–22, 2002.

Revolution Analytic, "Package ' doMC ,'" pp. 1–4, R package version 1.3.3, 2014.

Revolution Analytic and S. Weston, "foreach: Foreach looping construct for R," pp. 1-10, R package version 1.4.2, 2014.

W. R. Rudnicki, M. Kierczak, J. Koronacki, and J. Komorowski, "A statistical method for determining importance of variables in an information system," Rough Sets Curr., pp. 557–566, 2006.

V. Svetnik, A. Liaw, and C. Tong, "Variable selection in random forest with application to quantitative structure-activity relationship," Proc. 7th Course Ensemble Methods Learn. Mach., 2004.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "Misc functions of the Department of Statistics (e1071), TU Wien," R Packag., pp. 1–5, 2008.

L. Scrucca, "GA: a package for genetic algorithms in R," J. Stat. Softw., vol. 53, no. 4, pp. 1–37, 2012.