BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **21**, No 2
Sofia • 2021 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2021-0016

A New Noisy Random Forest Based Method for Feature Selection

Yassine Akhiat[1], *Youness Manzali*[1], *Mohamed Chahhou*[2], *Ahmed Zinedine*[1]

[1]*Faculty of Sciences, USMBA, Fez, Morocco*
[2]*Faculty of Sciences, UAE, Tetouan, Morocco*
E-mails: yassin.akhiat@gmail.com younes.manzali@usmba.ac.ma mchahhou@hotmail.com ahmedzinedine@yahoo.com

Abstract: *Feature selection is an essential pre-processing step in data mining. It aims at identifying the highly predictive feature subset out of a large set of candidate features. Several approaches for feature selection have been proposed in the literature. Random Forests* (*RF*) *are among the most used machine learning algorithms not just for their excellent prediction accuracy but also for their ability to select informative variables with their associated variable importance measures. Sometimes RF model over-fits on noisy features, which lead to choosing the noisy features as the informative variables and eliminating the significant ones. Whereas, eliminating and preventing those noisy features first, the low ranked features may become more important. In this study we propose a new variant of RF that provides unbiased variable selection where a noisy feature trick is used to address this problem. First, we add a noisy feature to a dataset. Second, the noisy feature is used as a stopping criterion. If the noisy feature is selected as the best splitting feature, then we stop the creation process because at this level, the model starts to over-fit on the noisy features. Finally, the best subset of features is selected out of the best-ranked feature regarding the Geni impurity of this new variant of RF. To test the validity and the effectiveness of the proposed method, we compare it with RF variable importance measure using eleven benchmarking datasets.*

Keywords: *Feature selection, data mining, random forest, Geni impurity, variable importance.*

1. Introduction

Due to the massive increase in data amount in real-world datasets, Feature Selection (FS) becomes a necessary pre-processing technique to reduce dimensionality. FS is the process of choosing relevant features and removing redundant, irrelevant and noisy ones [1-3, 25]. Generally, Feature selection aims to:

- Make models easier to interpret.

- Reduce resources requirement (shorter training time, small storage capacity, etc.).
- Avoid the curse of dimensionality.
- Avoid over-fitting problem, thus, better model generalization.
- Improve accuracy: less noise in data means improved modeling accuracy.

Feature selection can be categorized into three main approaches: 1) filters; 2) wrappers; 3) embedded approaches [4, 5, 25].

1. **Filter Methods** rely on the relationship between features and the class label (such as distance, dependency, correlation, etc.) to compute the importance of features. This category is a pre-processing step, which is independent from the induction algorithm. Filters are known by their ease of use and low computational cost. Let us mention some filter methods: Fisher score, Relief, Mutual information, Pearson correlation, and information gain-based methods, to mention a few [6, 7].
2. **Wrapper approach** generates models with subsets of features. Then, it uses prediction performance as a criterion function to lead the search for the best feature subset. This approach takes into account the interactions between features as opposed to Filters. Generally, Wrappers achieve a better performance than some Filter methods [5]. Wrapper methods include forward selection, backward elimination, and stepwise selection [5, 8-10].
3. **Embedded approach** performs feature selection implicitly while simultaneously constructing a model, which makes it less costly in terms of execution time than Wrappers. The mainly used embedded methods are the following: pruning methods, sparse learning-based methods such as L1 penalty and L2 penalty [5, 9, 10].

This paper mainly tackles the random forest for feature selection. The major contributions of this study are the following:

- **Noisy Random Forest (NRF).** We propose a new variant of Random Forest (RF) by adding a new stopping criterion to RF model. First, we add a noisy variable to a dataset; then, during the construction of the tree, if the noisy variable is selected as the best split feature, the construction process should be stopped. This step is meant to ensure the avoidance of the correlated features and the stability of the feature importance.
- **Feature ranking.** Based on the reliable feature importance of the proposed Noisy random forest, we rank features in a decreasingly reversed order. This step is reinforced to prevent choosing noisy and un-informative features. Moreover, the elimination is embedded by implication during the training of NRF.
- **Feature selection.** The explanatory ranked features in feature ranking step are used to construct a sequence of RF models by following a stepwise strategy. Then, the features of the last RF model are selected as the best subset.

Before sinking deep into details, let us put more emphasis on the highly relevant topics of this study (random forest, variable importance, feature selection).

1.1. Random forest

Random forest is a robust algorithm in different applications [11]. Many researches appreciate RF for their ability to handle the interaction between features, and they can be able to select informative features, especially in expression data analysis [12].

Based on the aggregation technique [13], RF combines several individual classifications or regression trees. Several bootstrap samples are drawn from the training data; then, a set of un-pruned decision trees are constructed on each bootstrap sample, so all trees of the forest are maximal trees. For each tree, a random subset of explanatory variables is selected for each split, and the best split is calculated only within this subset. From the fully constructed forest, the predicted class is obtained as the average of a majority vote of the prediction of all trees in the forest. The estimated prediction error of each tree is obtained using what is called the Out Of Bag samples (OOB), which is a set of observations that is not used for building the trees. Random forest is much stabler and accurate as compared to individual trees. Since RF is based on ensemble technique, it adjusts the instability that comes from the small changes in the learning sample [14]. The following steps give a more precise explanation of RF:

Step 1. Create a bootstrapped dataset: we randomly select samples from the original data set.

Step 2. Creating maximum decision trees (without pruning): from the created bootstrapped datasets, we build decision tree using just a random subset of variables at each step.

Step 3. Build a forest by repeating Step 1 and Step 2 for *N* times (*N* decision trees).

Step 4. Predicting the outcome: from the constructed forest, the prediction is obtained as an average or majority vote of the predictions of all trees.

Step 5. Evaluate the model: The prediction error is estimated using the set of observations, which are not used for building the current tree (called OOB).

1.2. Feature importance

Variable Importance (VI) measures of RF have received a lingering momentum in many applied tasks not only at the level of sorting features before a stepwise estimation model, but also in the trend of understanding and interpreting data. The basic variable importance of RF is the mean selection frequencies [14]. It counts the number of times each feature is selected in all trees. The most selected variable is the most important one. Another widely used VI is the Geni index, which measures how well a split on each variable is separating the samples of the two classes in this given node averaged over all trees [14, 15, 20]. These two indices are biased and not reliable when features are different in their scales, and when datasets contain many categorical features or noisy ones [14]. The most advanced variable importance of RF is “Mean decrease accuracy” [14, 16]. This measure is computed when data are permuted in OOB samples: RF importance variable is the difference between the prediction error recorded on out-of-bag samples and the prediction error after permuting the values of data averaged over all trees in the forest. The effect of the scale of measurement and number of categories on mean decrease accuracy is lesser than the effect on the mean selection frequencies and Geni index, but still hugely affects the reliability and interpretability of the variable importance measure.

1.3. Feature selection

Various feature selection algorithms based on the variable importance of RF have been introduced in the literature. Let us briefly mention some wrapper and embedded method based on variable importance:

1. The first wrapper methods based on VI coming from Classification And Regression Tree method (CART), see [11] and of course, random forests [15].

2. An algorithm is proposed in [17] to select useful variables using a stepwise strategy involving the CART. Based on Support Vector Machine (SVM) scores and relying on descending elimination.

3. Authors of [18] propose a new feature selection method based on Recursive Feature Elimination based Support Vector Machine (SVM-RFE) to evaluate variable subset relevance with regard to variable selection.

4. Another approach is presented in [19]. Relying on the Out of bag (OOB-error), it computes variable importance without recalculation at each step as [21]. Then, after fitting all RF models, the best-chosen solution is the model whose error rate is within U standard error of the minimum error rate of all forests.

5. Two step algorithm based on random forest importance in proposed in [16]. In the first step, variables ranked in a descending order are meant to identify explanatory variables highly related to the target variable. Then, variables of the smallest importance are to be removed. The chosen variables selected through the first step might be correlated and redundant. The objective of the second step is to select a small number of variables to achieve better accuracy. First, a collection of RF models is constructed using the best variables. The variables leading to the smallest error on OOB samples are selected. Second, a stepwise technique is used to build an ascending sequence of RF. Finally, the variables of the last model are chosen.

6. A Guided Regularized Random Forest (GGRF) is proposed in [22], where RF model is a model on the whole training set then, they utilize the feature importance to guide the feature selection process. In this method, the constructed trees may have high variance. In order to fix the previous problem in GGRF, [23] propose a Guided Random Forest (GRF) where each tree in GRF is constructed independently from any other.

Two different objectives for variable selection should remain quintessentially fundamental and deep-seated. First, it is of paramount importance to detect the significant features highly related to the response variable. Second, it is highly recommended to select a small subset of variables sufficient to construct an excellent parsimonious prediction of the response variable.

1.4. Outline

The paper is organized as follows. In Section 2, we illustrate a random forest feature importance technique, especially in the presence of various noisy features. Section 3 proposes a feature selection algorithm based on the proposed NRF stopping criteria. Section 4 examines some experimental results by focusing mainly on standard and high dimensional classification datasets. Finally, Section 5 opens a discussion about future work.

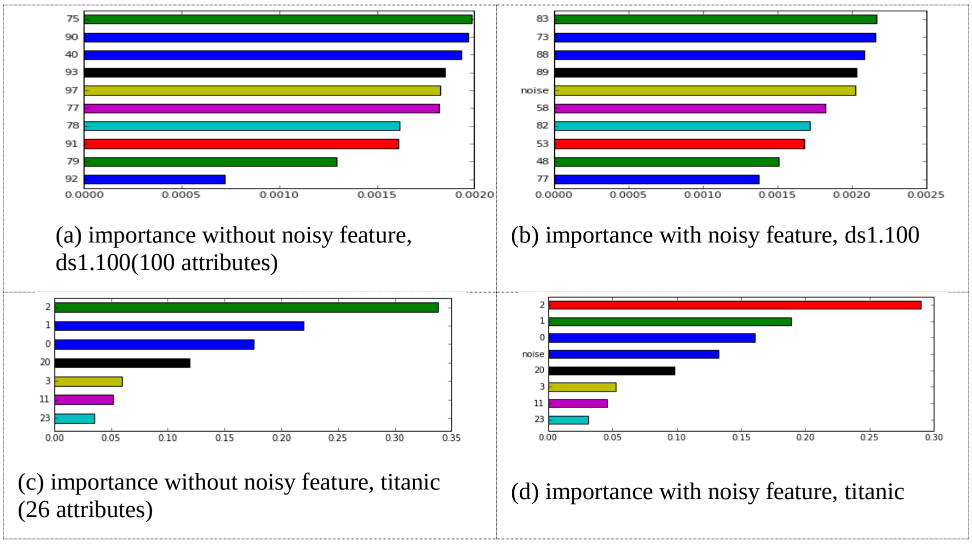
2. Motivation

Feature importance measures of random forest are among the widely used criteria as a means of variable selection in many classification tasks. RF importance (Geni impurity, etc.) computes the average decrease in contamination over all trees in the forest due to each feature. Removing the low ranked features according to their importance is not always a practical alternative because the ranking depends on the complexity of the model (tuning parameters). Some low ranked features can be more useful and informative if the complexity of the model is increased and vice versa.

To both illustrate and give more details about this behavior, we have conducted a simple experiment using two well-known datasets in the feature selection field: ds1.100 (100 variables) and titanic (26 variables). We have sorted features in descending order of RF importance using two scenarios:

1. The **First scenario.** We rank the variables in decreasing order using the original features of the two datasets.

2. The **Second scenario.** We rank the variables after adding a generated noisy feature to each dataset. The noisy feature is generated using a normal distribution. This choice based on the conducted experiment in (First experiment in Section 4).



(a) importance without noisy feature, ds1.100(100 attributes)

(b) importance with noisy feature, ds1.100

(c) importance without noisy feature, titanic (26 attributes)

(d) importance with noisy feature, titanic

Fig. 1. Ranking features in decreasing order according to RF Geni importance with and without noisy trick

As it is explicated through the illustrative figure (Fig. 1) below, the noisy feature (noise) is ranked as the best fifth feature among 100 features using the ds1.100 dataset (top right plot). On the titanic dataset, the noisy feature is ranked as the best fourth feature (bottom right plot) although the noisy feature is just a random noise. This is a conclusive empirical research-based assertion that exposes the impracticality of the pre-applied tendency. If we remove the lowest-ranked features first, as it is traditionally implemented by [11, 15, 16], we will probably forget the noisy feature because it is indiscernibly classified among the highly-ranked features. The RF model is probably overfitting on the noisy features, and by avoiding these noisy features during the moment of constructing the RF model first, some low ranked-features may become more useful to distinguish between classes. Thus, because of the reliable effectiveness of the practical proposed ranking-method mentioned above, there is a possible feasibility of avoiding the selection of noisy features as the most informative ones.

3. Proposed method

3.1. Procedure

To address the problem of unreliability of RF feature importance (Geni index) discussed in the previous section, we propose a feature selection method termed Noisy Random Forest (NRF). The algorithm consists of three main steps.

1. Proposing a new version of random forest called NRF:

- At each node, a noisy feature is added to the generated subset of features for the sake of splitting the current node.
- The noisy feature is used as a stopping criterion of RF.

2. Feature ranking:

- We sort features in decreasing order in accordance with NRF reliable importance.
- Feature elimination is performed by implication during the selfsame moment of the training phase. All features that are classified below the noisy feature are discarded. Denote by *k* the remaining features.
- Feature ranking step allows the selection of more features than necessary in order to make a careful choice later in the next step (feature selection).

3. Feature selection:

- A stepwise strategy is used to repeatedly construct a sequence of random forest models.
- Assess the AUC score of the model of the forest at each iteration.
- Reject a fraction of the least important features.
- The features of the last model are selected as the best subset.

The following three sub-sections provide an in-depth explanation and discussion of each NRF steps.

3.1.1. Noisy random forest

We propose a new stopping criterion of RF. First, we add a noisy feature (named noise) to the generated subset of features at each node. Then, this noisy feature is used as a stopping criterion as illustrated in this context: During the construction of the tree, in case the noisy feature is selected as the best splitting feature over other eligible ones, we stop the splitting process, and we return the internal node as a leaf node. The proposed criterion supplements a self-evident clarification, which guarantees that other features cannot be useful in splitting the data any further.

Fig. 2 and Algorithm NRF provide an in-depth explication of the idea.

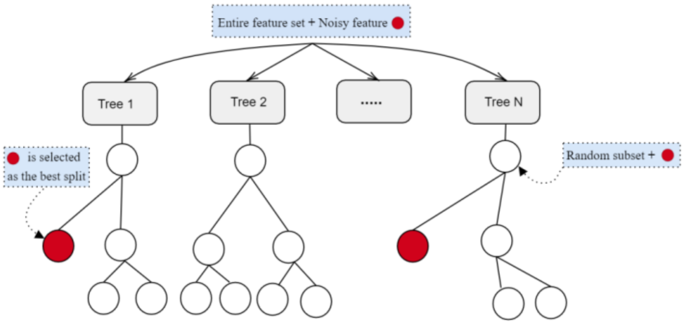

Entire feature set + Noisy feature ●
Tree 1
Tree 2
.....
Tree N
● is selected as the best split
Random subset + ●

Fig. 2. NRF procedure: NRF trees are expanded as RF until the noisy feature is selected as the best split. Then, the splitting process stops in the current branch

Why noisy feature is the stopping criterion?

In RF model, each tree is constructed as follows: at each node, a subset of features *S* is randomly generated to split the current node. The splitting feature is the one with the maximum Information Gain (IG). Assuming that the generated *S* contains just noisy and useless features, RF will compute IG of all features in *S* and it will choose the best splitting feature. Then, it will construct the tree without pruning. As a result, the constructed branches are complex and ineffective (of bad quality). This problem has motivated us to propose NRF, where a noisy feature is added to *S*. Thus, if the noisy (Red circle in Fig. 2) is selected as the best splitting feature, this means that all features in S are useless and they should not be included in the current tree and the splitting process is stopped. Using NRF, the constructed branches are short and consistent which may lead to construct simple and interpretable trees. In addition, NRF could hugely reduce the computational cost. Instead of computing the IG of all features at each node and constructing the full trees, NRF avoids the construction of sub-trees where the noisy feature is selected as the best split. The complexity of NRF in the worst case would be equal to RF (the case where Noisy feature is not selected which means the splitting process will continue and trees will be fully constructed without pruning as in RF) (see the conducted experiment in Table 1).

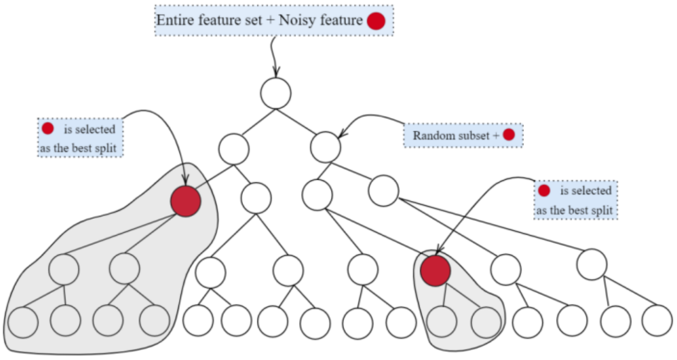

Entire feature set + Noisy feature ●
● is selected as the best split
Random subset + ●
● is selected as the best split

Fig. 3. The splitting process in RF and NRF. For RF, trees are fully constructed without pruning. In contrast to RF, NRF constructs a pruned trees using Noisy feature (Red circle). Once the Noisy feature is selected over other candidate features, the splitting process stops

The modified algorithm of random forest is the following:

Algorithm NRF

Pre-condition: A training set $\boldsymbol{S}:= (x_1; y_1),\ldots, (x_n; y_n)$, features **F**∪{**noisy feature**}, and number of trees in forest **B**

Output: A forest of trees.

```
Step 1.  function CreateTreesForest(S, F∪{noisy feature} )
Step 2.      H ← Ø
Step 3.      for $i \in$ 1,…, B do
Step 4.          $h_i$ ← RandomizedTreeLearn(S, F∪{noisy feature})
Step 5.          H ← H ∪ { $h_i$ }
Step 6.      end for
Step 7.      return H
Step 8.  end function
Step 9.      function randomizedTreeLearn(S, F∪{noisy feature})
Step 10.     At each node:
Step 11.         f ← a subset of F
Step 12.  best_split←choose the best split feature from f∪{noisy feature}
Step 13.             If best_split = noisy feature then
Step 14.  return the learned tree
Step 15.             Else
Step 16.                     split on best_split
Step 17.     return the learned tree
Step 18.       end function
```

3.1.2. Feature ranking

We sort features in decreasing order of NRF importance. As opposed to [14, 16, 18, 22], our approach does not need variable elimination since it eliminates unimportant and noisy features by implication during the learning process. In the field of feature selection, it is an intelligibly well-known fact that variables with high redundancy might be present in any datasets. Thus, the NRF model can use any of these correlated features. Once one of these correlated features is used as a predictor, the importance of others is exponentially decreased because the impurity, which the correlated features can decrease, is already reduced by the first used feature. Therefore, they will be quantified of below-average and inconsequential importance. As a result of what has been articulated above, the ranked features of the proposed method are not correlated or redundant as is in [16, 18, 22].

3.1.3. Feature selection

From the subset of the best-ranked features selected in the second step, a necessary attempt should be made to find a small number of features applicable to an excellent parsimonious prediction of the response variable. The stepwise technique is applied when RF models are repeatedly constructed and the worst features are discarded until the examination of all features. The search strategy is guided by the grid search strategy and the AUC score. Thus, the features of the last best performing model are selected.

3.2. Complexity analysis

- Training phase

Time complexity is the number of required operations for building models based on data. Time complexity of RF is $O(Bmn\log(n))$ where B is the number of constructed trees, m is the number of features to sample at each node and n is the number of data samples [26]. This is the worst-case scenario since RF trees are fully constructed, which means nodes are expanded until all leaves are pure (depth=None).

For our NRF model there is always a possibility for nodes to stop expanding, if the "Noise" is selected as the best split (see Algorithm NRF, Step 13). This advocates the fact that in the worst cases, the complexity of NRF would be equal to RF complexity; otherwise, NRF complexity is always less than the one of RF.

- Selection phase

Our suggested method consists of two main steps. The first one is the common feature ranking where features are classified in accordance to their importance. Since our NRF does not allow noisy feature to be included in RF branches, the insignificant features are eliminated by implication during the training phase (as demonstrated in Experiment 2 and Experiment 3). As a result, the burden of eliminating the un-informative and redundant features has been already avoided. In the second step of NRF, we put more emphasis on finding a small number of features applicable to an excellent parsimonious prediction of the response variable.

Table 1.The execution time (in seconds) for both RF and our NRF versus the number of times the noisy feature is selected as the best split.

Dataset	RF	NRF	Number of noisy feature
Sonar	30	**29**	7
Chess	47.92	**31**	79
Spambase	224	**218**	20

To provide more information about the proposed feature selection procedure, the following example is suggested.

3.3. Starting example

For further explanation and illustration of the proposed method, we apply the feature selection procedure on the clean dataset, which is a binary classification dataset of 167 attributes and 6600 instances.

Training NRF. We train the new version of the random model NRF, where a noisy feature is used as a stopping criterion. Training shorter trees can be a practical alternative as espoused by the fact that the representative features always appear in the few first levels [17]. For this reason, the parameters used in NRF are Depth =3 and number of trees Ntree= 100.

Feature ranking. After training NRF, features are ranked in a decreasing order according to their NRF importance. All features that are ranked below the noisy one are to be discarded.

The result displayed in the Fig. 4 shows the ranking of features according to their importance. The high-ranked features are more important than noisy feature.

We keep only the features whose importance highly exceed and outperform the noisy feature's poor one. This step leads to retaining more features than necessary. (For the clean dataset, the selected features in the Step1 is *k*=19).

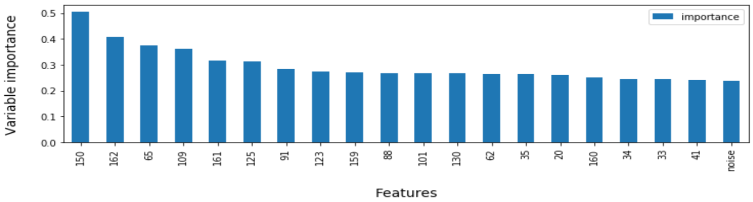


Fig. 4. Feature importance using the proposed NRF (clean dataset)

Feature selection. From the *K* best features, we repeatedly construct a random forest models with a grid search strategy and remove the underperformed feature (lower AUC score). Then, we make a consecutive repetition of this process with the remaining features until all of them are examined in terms of performance. The features of the last model are selected.

The following graph shows the results of the feature selection step (Step 2). Note that the AUC score increases quickly and reaches its maximum when the first 17 informative features are included in the model (the AUC score is higher that 96%). Then it remains nearly constant. This means that the best subset contains the first 17 attributes.

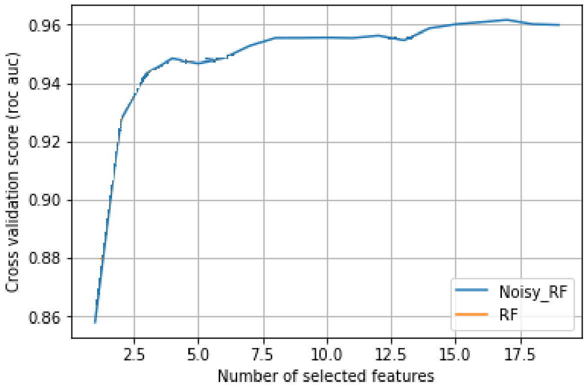


Fig. 5. The performance of the feature selection procedure for clean dataset

4. Experimental results

The proposed method is compared to the standard RF approach in terms of prediction AUC score. Three experiments have been conducted to assess the validity of our method.

First experiment. This experiment is conducted to empirically justify and substantiate the choice of normal distribution over other distributions.

Second experiment. All variables are chosen to be equally irrelevant to scrutinize the stability and the reliability of the proposed method. The reliable variable importance should not prioritize any predictor variable over other.

Third experiment. This simulation is conducted to evaluate the ability of the proposed algorithm to deal with correlated variables.

Fourth experiment. This experiment is performed to appraise the performance of the NRF in terms of AUC score using ten standard benchmarking datasets and one high dimensional classification one.

4.1. Datasets

Through this manuscript, we employed ten binary classification datasets to experiment the performance of the proposed feature selection method. All datasets can be downloaded from UCI machine leaning repository [24] and Kaggle platform. Datasets are chosen to be different in terms of attributes and instances to validate the efficiency of the proposed algorithm. The characteristics of each dataset are summarized in the following table:

Table 2. Characteristics of the benchmarking datasets

Type of dataset	Dataset	Features	Instances	Distribution	Class
Standard datasets	ds1.100	100	26,733	3% + / 97% –	2
	credit card	24	30,000	22% + / 78% –	2
	ionosphere	34	351	64% + / 36% –	2
	spambase	57	4601	39% + / 61% –	2
	Musk	167	6598	15% + / 85% –	2
	chess	36	3196	52% +/ 48% –	2
	caravan	86	5822	6% + / 94% –	2
	madelon	500	4400	50% + / 50% –	2
	eighther	306	200,000	4% + / 96% –	2
	sonar	59	208	47% + / 53% –	2
High dimensional datasets	colon	2000	62	65% + / 35% –	2

4.2. Results and discussion

Experiment 1. Why we choose normal distribution. The used noisy feature in NRF could be generated using different distributions (Normal, geometric, exponential, etc.). In the experiments conducted in this manuscript, we choose to use the normal distribution rather than other distributions regarding to its good empirical results in terms of AUC score as illustrated in Table 3.

Table 3.The impact of different distributions on the performance of NRF in terms of AUC score and execution time(in seconds)

Dataset	Normal distribution		Exponential distribution		Geometric distribution		Binomial distribution	
	Time	AUC	Time	AUC	Time	AUC	Time	AUC
Sonar	30	**92%**	33.15	91.5%	**29.61**	**92%**	31.20	91.3%
Chess	47.92	**99%**	44.95	94%	5.10	94%	**4.87**	95%
Spambase	224	**97.8%**	206.16	96%	**138.25**	97%	139.67	96.9%
Means	100.64	**96.3%**	94.75	93.8%	**57.65**	94.3%	58.58	94.3%

Experiment 2. All features are equally irrelevant. In this simulation, when all features are equally not useful and irrelevant, the variable importance of the RF and the proposed NRF (Noisy_RF) are supposed to be the same. However, as it is illustrated and presented in Fig. 6 (top plot), the importance of variables is considerably different from one variable to another. As opposed to RF variable importance, the variables' importance of the proposed variant of RF (Noisy_RF) are equally presented and there is no preference of any variable over the others (bottom plot). Accordingly, all of them are deemed to be irrelevant, and therefore they should be discarded. Thus, the drawn conclusion stemmed from the reached implications espouses the following experiment-based assertion: unlike the Geni importance of RF, which cannot reliably measure the variable importance, our variable importance measure is dependable and unbiased.

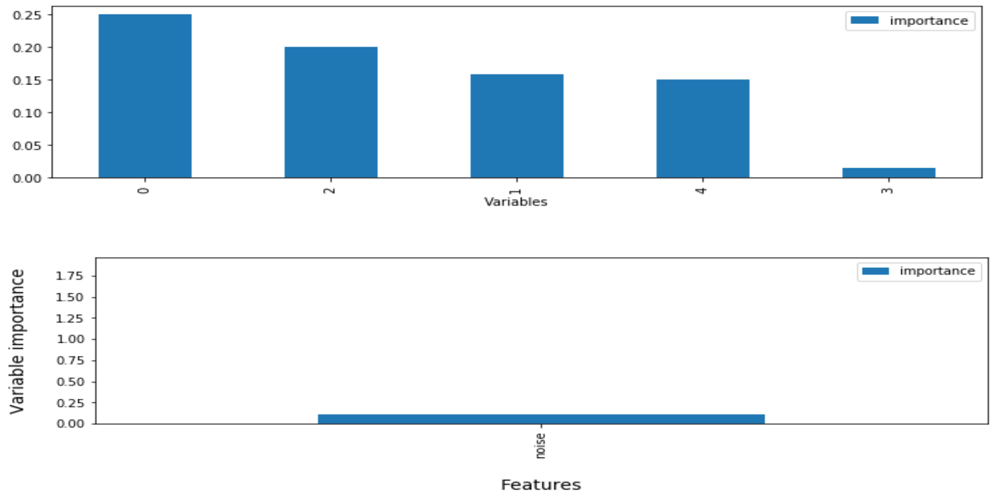


Fig. 6. Variable importance measured by RF and our NRF. The top plot displays the RF variable importance while the bottom one presents the NRF variable importance

Experiment 3. The presence of redundant and correlated features. To demonstrate the functionality of variable importance's behavior of the standard RF and the proposed variant of RF, an explicatory experiment is conducted in which we applied the proposed method (Noisy_RF) and RF variable importance on a generated dataset that contains correlated variables. The best variable measure is meant to disable the correlated and redundant features from being selected. The results show that the importance measured by the standard RF of all variables are equal (see the top plot) which may allow the redundant and correlated features to be opted as the best feature subset. This problem is tackled through the application of the proposed variable importance measure (bottom plot). Therefore, the selected subset is more consistent and diverse.

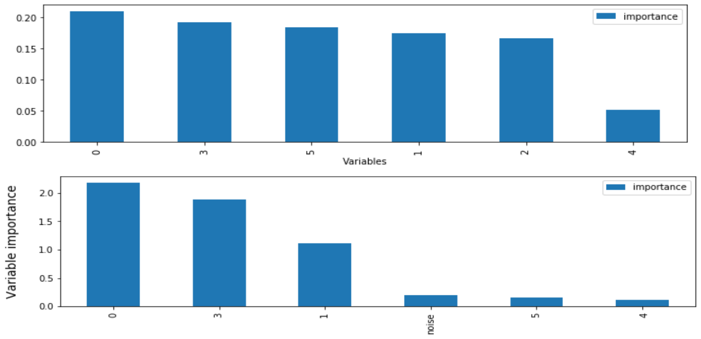


Fig. 7. Variable importance in the presence of correlated features. The top plot displays the RF variable importance while the bottom one presents the NRF variable importance

Experiment 4. The performance of Noisy_RF on ten benchmarking datasets. In this experimental study, we compared the performance of our feature selection procedure (Noisy_RF) to the RF. The comparison is carried out counting on the datasets shown in Table 1. The quality of the final selected subset for both methods is evaluated through the application of a 3-fold cross-validation to estimate the AUC score rate. So we split each dataset into three stratified folds, each fold is used as a test set, and the remaining folds are used as a training set. The results are obtained as those of Fig. 4, except that for the ranking feature step, we only plot the 50 most important variables to ensure the clarity and the apparent visibility of the graphs. We opted for the usage of the AUC metric because it is more convenient for the evaluation of classifiers performance on unbalanced datasets.

- **Standard datasets.** The results obtained for the caravan dataset using our procedure showed that the first step (feature ranking) enables the selection of 26 features only (Fig. 8 the top left plot). After the application of the feature selection step (Fig. 8 the right plot), it is notably conspicuous that with only ten features, the AUC score is in a cumulative growth as it has reached the percentage of 77.2 %. The unprecedented attainability of the AUC score remarkably displays the powerful performance of the 10 selected features as opposed to the number of the uninformative features discarded (88.4% of features are eliminated). On the other hand, the results of the Random forest corroborate the fact that despite the large number of the selected features in the first step, *k*=47 (Fig. 8 the bottom plot), which provides the potential likelihood of selecting the best features in the second step, the best performance achieved is exclusively restricted in the percentage of 76.1 %. This comparative study confirms that disregarding the powerfully relevant and informative features, the RF variable importance measure tends to select the unreliably biased noisy features as the high-ranked ones.

The results on ionosphere dataset demonstrated that the elimination step (Fig. 9 the top left plot) prompts the obliteration of 56 % of unimportant features using NRF and the removal of 6 % using RF. Then, from the remaining *K* features, seven features are selected as the reliable subset in Step 2 with the maximum AUC score of 97.5% (Fig. 9 the right plot) for our NRF. Whereas, the RF maximum performance is restrictively reduced to the percentage of 97.4%. Thus, no pervasive disparity can be discerned between the maximum AUC score for NRF and RF, yet the ammount of features selected by RF in first step is time and memory consuming.

The same NRF elimination procedure when applied on spambase dataset leads to the removal of more than 33 % of useless features. Out of 66 % of the remaining features, 79 % are selected from the last model with an AUC score of (97.8 %) (Fig. 10). RF has achieved the same results, yet our proposed method has slightly outperformed it with a further step manifested in the attainability of AUC score and the best-ranked features selection in the first elimination step.

For the clean dataset, which contains 166 features (Fig. 11), the procedure of feature ranking of NRF engenders the preservation of 19 features only. Whereas, RF leads to the selection of more than 100, which is a largely massive number compared to the pre-selected one of the preserved features. Relying on just the first 17 selected features of our method, the AUC score has reached its maximum. This considerable dimensionality reduction (about 90% of features are eliminated) enables the construction of fast models and decreases the storage and memory requirement.

The same results are obtained on datasets eighther, madelon, ticdata2000, credit card, and chess dataset (see Figs 12-16). The elimination step of our proposed method always selects the smallest, reliable and consistent feature subset compared to RF, which leads to the achievement of the highest AUC score in the feature selection step. This conducted experiment obviously confirms that features could reliably measure the importance of features by applying NRF even in situation where correlated and redundant features are present or when features are varied in their scale of measurement. Moreover, the performance of feature subset selected in the feature selection step drastically outperforms RF performance almost in all datasets in terms of AUC score attainability.

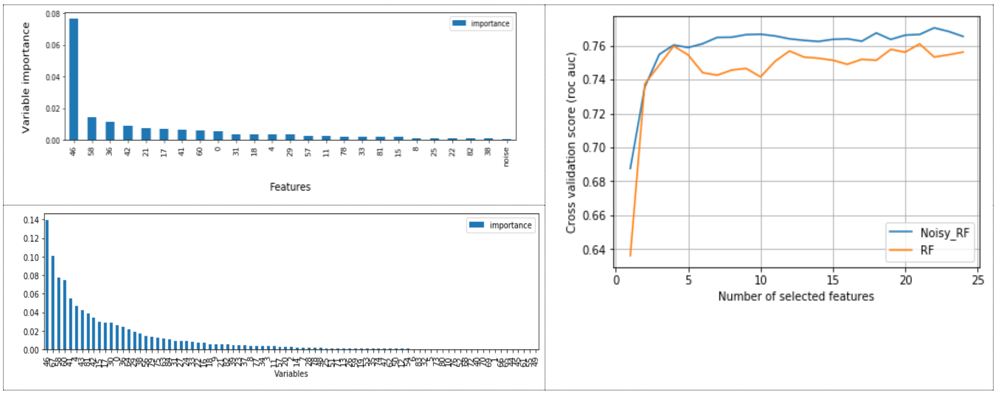


Fig. 8. Feature ranking and Feature selection applied on caravan dataset

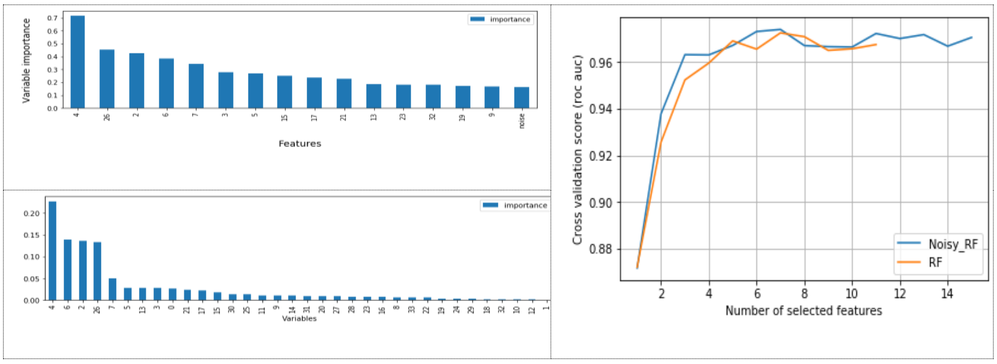


Fig. 9. Feature ranking and Feature selection applied on ionosphere dataset

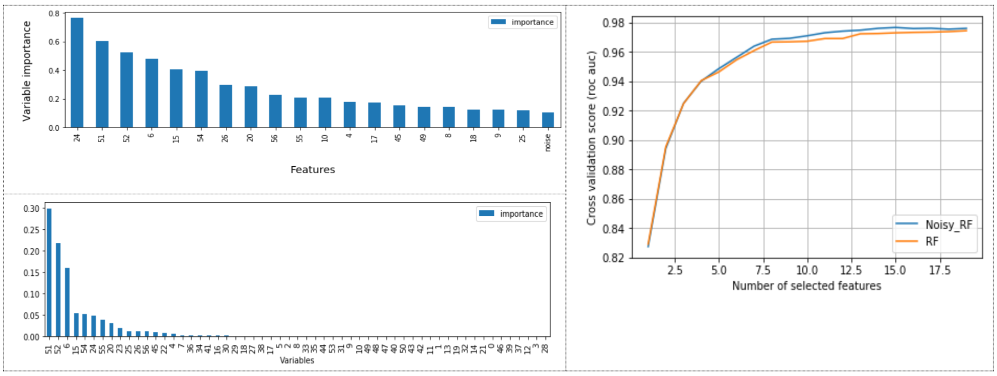


Fig. 10. Feature ranking and Feature selection applied for spambase dataset

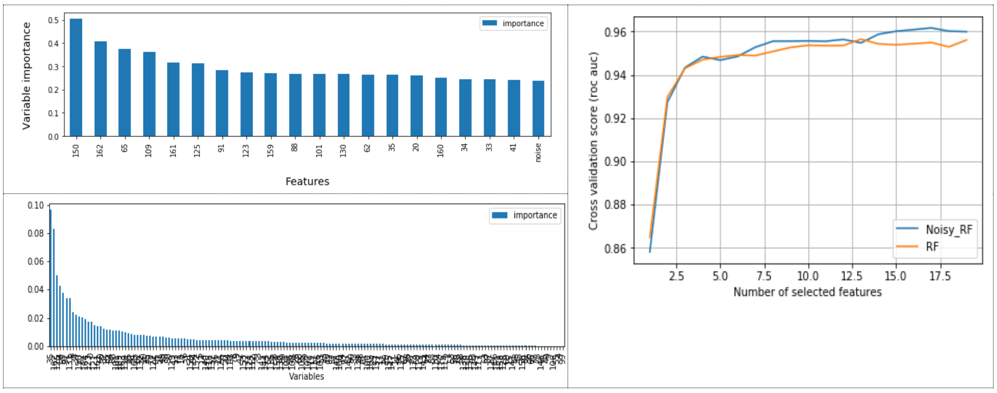


Fig. 11. Feature ranking and Feature selection applied for clean dataset

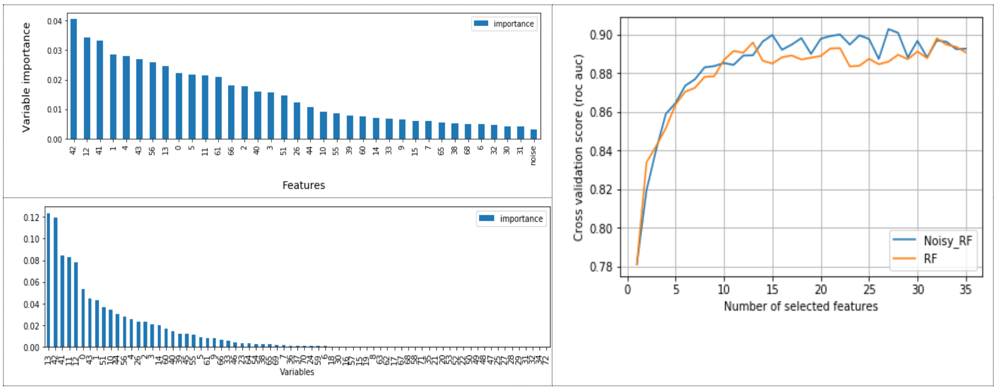


Fig. 12. Feature ranking and Feature selection applied for eighther dataset

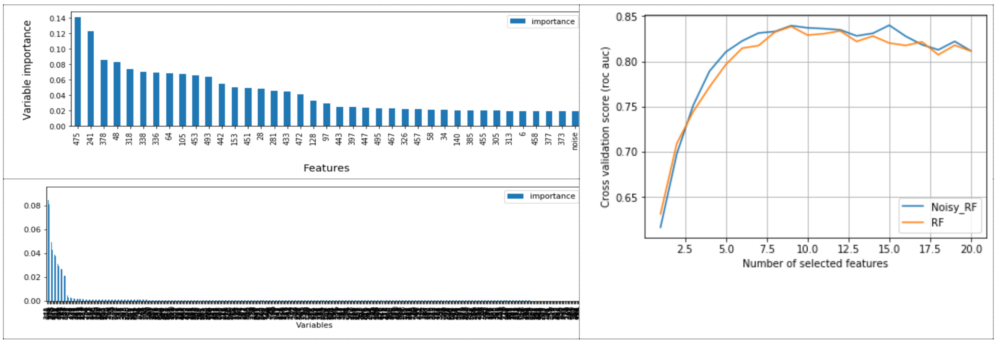


Fig. 13. Feature selection and feature selection for madelon dataset

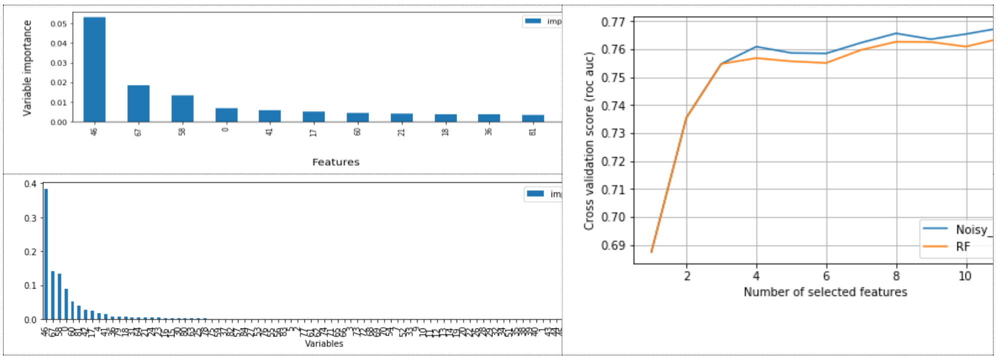


Fig. 14. Feature selection and feature selection for ticdata2000 dataset

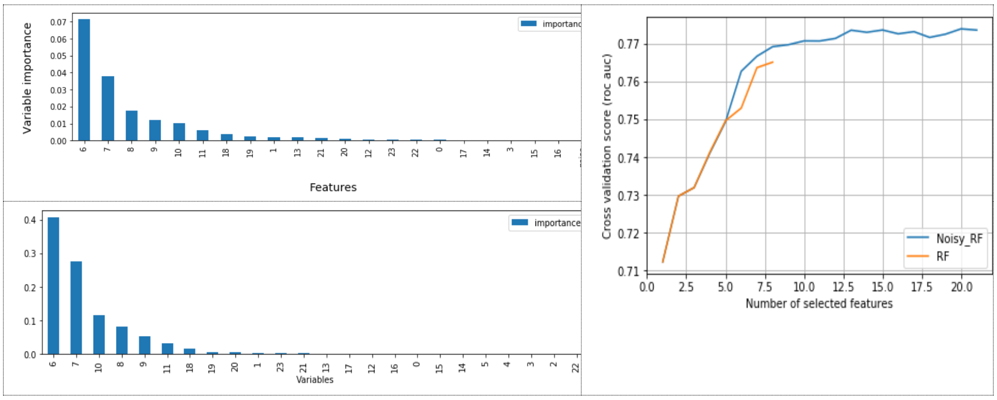


Fig. 15. Feature selection and feature selection for credit card dataset

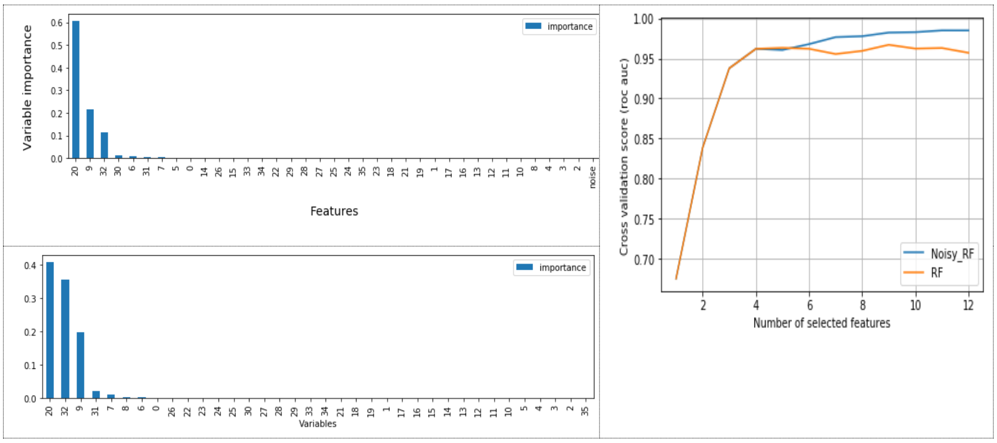


Fig. 16. Feature selection and feature selection for chess

- **High dimensional classification dataset.** The method introduced is also applied on the well-known high dimensional dataset called colon (2000 features and 62 examples) see Table 1 to estimate the prediction performance. Since these types of datasets are of small size, we used a 5-fold cross validation so that the training set can contain enough training examples. The drawn results (Fig. 17) on colon dataset accentuate that the proposed method has the ability to select the best unbiased feature subset even in extreme cases in which datasets contain high number of features or small number of examples.

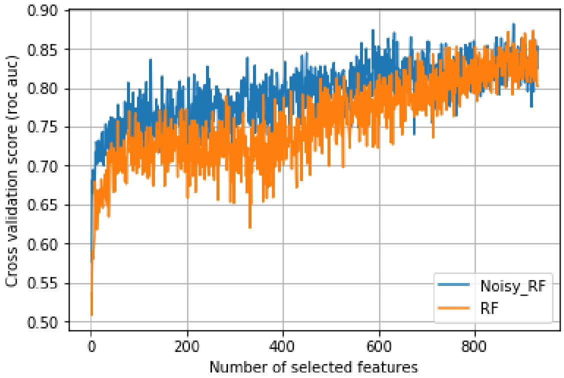


Fig. 17. Feature selection for colon dataset

5. Conclusion and future scope

Random forest introduced by Leo Breiman in 2001 is a powerful machine learning model that has been applied in many real word problems not just for its outstanding performance but also for its variable importance measures as means of feature selection. Geni impurity is a widely used variable importance measure. Through this paper, we have demonstrated empirically that this variable importance measure cannot be reliably applied in feature selection in situations when datasets contain huge amount of correlated, redundant and features of varying type and scale of measurement. Therefore, we have proposed an alternative variant of random forest that provides a reliable and unbiased feature importance measure as means of feature selection counting on the proposed noisy feature technique. Our NRF method has shown its ability to measure reliably the variable importance compared to RF. Moreover, it is capable of selecting the smallest consistent and diverse feature subset, which leads usually to better performance, minimum resources and storage requirement.

In the future works, we will consider the highly advanced variable importance measure, which is the mean decrease in accuracy since the effect of the scale of measurement and the number of irrelevant and correlated features has minor influence. Besides, we will examine and evaluate the new variant of random forest in classification and regression problems.

References

1. Akhiat, Y., M. Chahhou, A. Zinedine. Ensemble Feature Selection Algorithm. – International Journal of Intelligent Systems and Applications, Vol. **11**, 2019, No 1, p. 24.
2. Akhiat, Y., M. Chahhou, A. Zinedine. Feature Selection Based on Pairwise Evalution. – In: Proc. of 2017 Intelligent Systems and Computer Vision (ISCV'17), IEEE, 2017.
3. Akhiat, Y., M. Chahhou, A. Zinedine. Feature Selection Based on Graph Representation. – In: Proc. of 5th International Congress on Information Science and Technology (CiSt'18), IEEE, 2018.
4. Venkatesh, B., J. Anuradha. A Review of Feature Selection and Its Methods. – Cybernetics and Information Technologies, Vol. **19**, 2019, No 1, pp. 3-26.

5. Li, J., et al. Feature Selection: A Data Perspective. – ACM Computing Surveys (CSUR), Vol. **50**, 2017, No 6, pp. 1-45.
6. Urbanowicz, R. J., et al. Relief-Based Feature Selection: Introduction and Review. – Journal of Biomedical Informatics, Vol. **85**, 2018, pp. 189-203.
7. Gu, Q., Z. Li, J. Han. Generalized Fisher Score for Feature Selection. – arXiv preprint arXiv:1202.3725, 2012.
8. Huang, S. H. Supervised Feature Selection: A Tutorial. – Artif. Intell. Research, Vol. **4**, 2015, No 2, pp. 22-37.
9. Jović, A., K. Brkić, N. Bogunović. A Review of Feature Selection Methods with Applications. – In: Proc. of 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'15), IEEE, 2015.
10. Chandrashekar, G., F. Sahin. A Survey on Feature Selection Methods. – Computers & Electrical Engineering, Vol. **40**, 2014, No 1, pp. 16-28.
11. Breiman, L. Random Forests. – Machine Learning, Vol. **45**, 2001, No 1, pp. 5-32.
12. Díaz-Uriarte, R., S. A. de Andres. Gene Selection and Classification of Microarray Data Using Random Forest. – BMC Bioinformatics, Vol. **7**, 2006, No 1, p. 3.
13. Breiman, L. Bagging Predictors. – Machine Learning, Vol. **24**, 1996, No 2, pp. 123-140.
14. Strobl, C., et al. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. – BMC Bioinformatics, Vol. **8**, 2007, No 1, p. 25.
15. Breiman, L., et al. Classification and Regression Trees. CRC Press, 1984.
16. Genuer, R., J.-M. Poggi, C. Tuleau-Malot. Variable Selection Using Random Forests. – Pattern Recognition Letters, Vol. **31**, 2010, No 14, pp. 2225-2236.
17. Poggi, J. M., C. Tuleau. Classification supervis´ee en grande dimension. Application `a l'agr´ement de conduite automobile. – Revue de Statistique Appliqu´ee, LIV, Vol. **4**, 2006, pp. 39-58.
18. Rakotomamonjy, A. Variable Selection Using SVM-Based Criteria. – Journal of Machine Learning Research, Vol. **3**, March 2003, pp. 1357-1370.
19. Díaz-Uriarte, R., S. A. de Andres. Gene Selection and Classification of Microarray Data Using Random Forest. – BMC Bioinformatics, Vol. **7**, 2006, No 1, p. 3.
20. Menze, B. H., et al. A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. – BMC Bioinformatics, Vol. **10**, 2009, No 1, p. 213.
21. Jiang, H., et al. Joint Analysis of Two Microarray Gene-Expression Data Sets to Select Lung Adenocarcinoma Marker Genes. – BMC Bioinformatics, Vol. **5**, 2004, No 1 p. 81.
22. Deng, H., G. Runger. Gene Selection with Guided Regularized Random Forest. – Pattern Recognition, Vol. **46**, 2013, No 12, pp. 3483-3489.
23. Deng, H. Guided Random Forest in the RRF Package. – arXiv preprint arXiv:1306.0237, 2013.
24. Dua, D., C. Graff. UCI Machine Learning Repository. 2019. Irvine, CA, University of California, School of Information and Computer Science, 2003.
    http://archive.ics.uci.edu/ml
25. Venkatesh, B., J. Anuradha. A Review of Feature Selection and Its Methods. – Cybernetics and Information Technologies, Vol. **19**, 2019, No 1, pp. 3-26.
26. Gilles, L. Understanding Random Forests: From Theory to Practice. – arXiv preprint arXiv:1407.7502, 2014.

Received: 27.08.2020; Second Version: 12.01.2021; Accepted: 22.02.2021