

Reliable Attribute Selection Based on Random Forest (RASER)

Aboudi Noura^{1(✉)}, Hechmi Shili¹, and Lotfi Ben Romdhane²

¹ Department of Computer Science, University of Monastir, Monastir, Tunisia
aboudinoura89@gmail.com, shilihechmi@gmail.com

² Department of Computer Science, University of Sousse, Sousse, Tunisia
lotfi.ben.romdhane@gmail.com

Abstract. Feature selection has become one of the most active research areas in the field of data mining. It allows removing redundant and irrelevant data sets of large size. Furthermore, there are several methods in the literature for selecting attributes. In this article, a new multi-objective method is proposed to select relevant and non-redundant features. Our proposed feature selection method is divided into three stages: The first step computes the feature relevance value based on random forests. The second step, computes the dissimilarity matrix representing the dependence between the features of our training datasets, and transform it into a complete graph whose nodes represent features and edges represent the values of dissimilarities between them. The last step is for the optimization in which a multi-objective optimization algorithm is applied. The proposed method is applied on many datasets to find the most relevant and non-redundant features and the performance of the proposed method is compared with that of the popular MBEGA, mRMR (MIQ) and mRMR (MID).

Keywords: Feature selection · Feature relevance · Feature redundancy · Multi-objective optimization · Random forest

1 Introduction

Feature selection is a problem that has to be addressed in many areas. The resolution of general problems is based on the treatment of features [1]. So the performance of the treatment system depends on the correct selection of features: considering many features increases the risk of considering redundant and irrelevant features. Therefore, it is necessary to use a method for reducing the data size. The feature selection identify a subset of attributes of minimal size necessary and sufficient to define the target concept [2]. Generally, a typical feature selection method illustrated in Fig. 1, consists of four components: a subset generation or search procedure, an evaluation function, a stopping criterion, and a validation procedure [3]. The selection of attributes will be based on two criteria: relevance and redundancy.

Definition 1 Relevance: In the literature, there are several definitions of the concept of relevance of a feature, the best known is that of [4, 5]. A feature f_i is said relevant if its absence reduces significantly the performance of the used classification system.

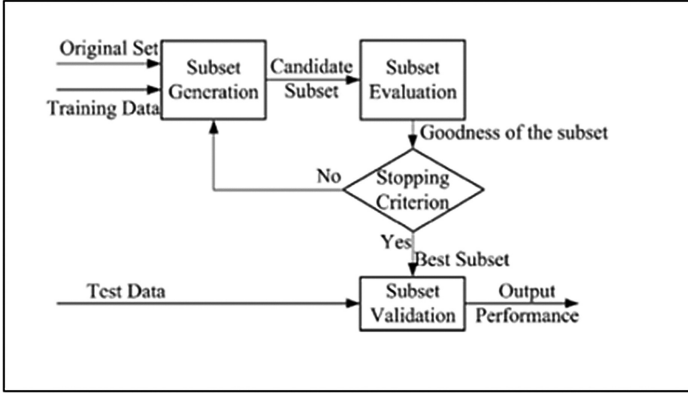


Fig. 1. General feature selection process [3]

Definition 2 Redundancy: The notion of redundancy is relatively associated with correlation function. So, we say that two attributes are redundant if their values are completely correlated [6]. The concept of redundancy will be defined with the Markov blanket [7].

If G be a set of attributes, an attribute is redundant and can be removed from G if it is not relevant and Markov blanket M in G . According to the property of Markov blanket, it is easy to see that a redundant attribute remains removed redundant when other attributes are removed.

2 Related Work

In the literature, feature selection methods are generally grouped into three categories: filter, wrapper and embedded approaches [8]. The filter approach uses statistical measures calculated to filter characteristics according to the number of criteria. This step is generally carried out before applying any classification algorithm [8]. In filter approach, there are several methods such as minimal and redundancy-maximal-relevance criterion (mRMR) [9]. The mRMR is a screening method for the selection of attributes who have the best interests of the target class, are a little redundant and most dissimilar to each other. This method is based on statistical measures such as mutual information, correlation criteria, etc. The two optimization criteria (maximum relevance (MR) and minimum redundancy (mR)) are based on mutual information [10]. There are two types of mRMR that is different depending on the combination of the two criteria (addition and : mRMR(MID) and mRMR(MIQ).

$$\text{mRMR(MID)} = \max(\text{Pertinence} - \text{Redundancy}) \quad (1)$$

$$\text{mRMR(MID)} = \max(\text{Pertinence} - \text{Redundancy}) \quad (2)$$

Besides, there are many other methods such as: Correlation based on feature selection (CFS), Markov blanket filter (MBF)...

The wrapper methods on the contrary, use induction algorithm to evaluate the candidate feature subsets. They generally select feature subsets more suitable for the induction algorithm than the filter methods. In wrapper approach, there are several methods such as: Markov blanket-embedded Genetic Algorithm (MBEGA) [3] provided by Zexuan and Al in the selection of genes. MBEGA used Markov blanket to narrow the search by adding some relevant attributes or removing redundant and/or irrelevant attributes in the solutions selected by genetic algorithm (GA).

The embedded feature selection method, similar to wrapper methods and the feature selection is linked to the classification stage. The embedded methods have been proposed to reduce the classification of learning. They try to combine the advantages of both previous methods. The learning algorithm takes advantage of its own variable selection algorithm. See for example: PLSRFE [11], RF-RFE [12].

Most of the traditional feature selection methods define some metrics to evaluate each individual feature, such as signal-to-noise ratio (SNR) [6] and information gain [13]. They also use many statistical hypotheses testing techniques such as parametric t-test, F-test [11]. Since evaluating 2^N subsets becomes an NP-hard problem, suboptimal subsets are found by using search algorithms that find a subset heuristically. A number of search algorithms can be used to find a subset of variables that maximizes the objective function which is the classification performance [14].

3 Reliable Attribute Selection Based on Random Forest (RASER)

The feature selection is applied to reduce the number of attributes in many applications. Most features selection methods mainly focus on finding the relevant features [15]. We show that the notion of relevance is not sufficient for effective selection of high-dimensional data because of the high correlation that may exist between different features.

In many problems, there exist different aspects of solutions which are partially or wholly in conflict. Therefore, treating these problems as single objective optimization produces an unreliable result. In multi-objective optimization problem the objectives may estimate those different aspects of solutions which are conflicting in nature. After defining the problem of multi-objective optimization, meta heuristics are designed to solve the multi-objective problems. The goal of a multi-objective optimization is to identify a set of solutions in the Pareto optimal set [6].

This section is devoted to the explanation of our contribution, which is to propose a new method for the selection of relevant non-redundant features on the basis of the random forests method. Our method is divided into three steps. The approach followed is recapitulated in Fig. 2.

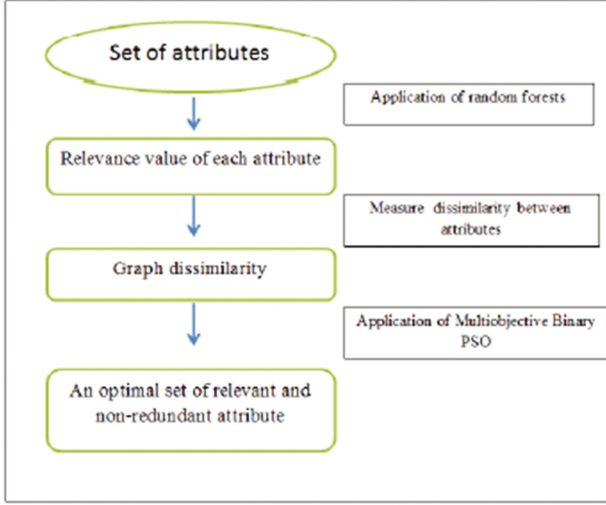


Fig. 2. The steps of our model

First, we compute the relevance values of each feature using random forests. In the second step, we compute the dissimilarity matrix from our training set by using the correlation criterion. Then, our problem is represented by a complete graph whose nodes represent the attributes and the edges represent the value of dissimilarity between them.

The last step is the optimization step in which we apply a multi-objective optimization algorithm (relevance, redundancy) on the dissimilarity graph to obtain an optimal subgraph embedding the most relevant non-redundant features.

3.1 Measuring Features Relevance

Random forests do not require the reduction of the prediction space before classification. In addition, random forests measure the relevance of features for each predictor candidate [16]. This study examines the effectiveness of variable importance measures by random forests in identifying the true indicator of a large number of candidate predictors. Random forests provide an original way for calculating the relevance value of feature. Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single feature. So, in a random forest or each node of the tree represents an attribute. Random forests handle the problem of relevance but not the problem of redundancy. Therefore, we will use the graph dissimilarity to handle features redundancy. Algorithm 1 is used for computing of the relevance value [16].

Algorithm 1. Measure of relevance value using random forests

```

1 Input:  $X_0 = [x_1, x_2, x_3, \dots, x_k]^T$  :training samples
2  $Y = [y_1, y_2, y_3, \dots, y_k]^T$  : classes labels
3 Output:list of attribute relevance values  $R$ 
4 Begin
5   initialization: $N_i = 0, M_i = 0, M_i^j = 0$ ,
6    $N_i$ = Number of times when the observation  $i$  appears in the OOB samples.
7    $M_i$ =Number of times when the observation  $i$  appears in the OOB samples, and misfiled.
8    $M_i^j$ = Number of times when the observation  $i$  appears in the OOB samples, and misfiled
   after permutation of values of the variable  $j$  in OOB.
9   For  $i = 1..n$  do
10     For  $j = 1..p$  do
11       For  $k = 1..K$  do
12         IF (observation  $i$  is in  $OOB_k$ ) THEN
13            $N_i = N_i + 1$ 
14         End IF
15         IF (observation  $i$  is in  $OOB_k$  and misclassified) THEN
16            $M_i = M_i + 1$ 
17         End IF
18         Swap the values of the random variable  $j$  in  $OOB_k$ .
19         IF (observation  $i$  is in  $OOB_k$  and misfiled after swapping the variable  $j$  values in
           OOB) THEN
20            $M_i^j = M_i^j + 1$ 
21         End IF
22       End For
23     End For
24   End For
25
26   End For
27   the relevance of the  $j$  variable =  $\frac{1}{n} \sum_{i=1}^n Z_i(j)$  where  $Z_i(j) = (M_i^j - M_i)/N_i$ 
28 End

```

3.2 Measuring Redundancy

The feature selection process is represented by the graphs, whether $G = (V; E)$ a graph constructed by a set of nodes V and a set of edges $E \subseteq V \times V$; and W is a weight matrix whose values are in the range $[0:1]$. Each node of the graph represents an attribute and each edge between two attributes represents the relationship between them.

For a set A of N attributes where $A = \{a_1; a_2; a_3; \dots; a_N\}$, the arrangement of the training set may be considered as a two-dimensional matrix where the columns represent the attributes and rows represent instances attributes.

There are a variety of methods for measuring of similarity dissimilarity between the attributes such as: Euclidean distance, correlation coefficient [6], etc. By using one of these dissimilarity measures a symmetric matrix is generated called dissimilarity matrix. Therefore if we have a set of N attributes $A = \{a_1; a_2; a_3; \dots; a_N\}$, computation of dissimilarity between attributes is handled by a S_m .

The dissimilarity matrix is computed from the data matrix by using the correlation coefficient [1] between each pair of attributes. The correlation coefficient σ between two random attributes x and y is defined by:

$$\sigma(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}} \quad (3)$$

Where $var(.)$ denotes the variance of a variable and $cov(.)$ the covariance between two variables. If x and y are completely correlated (i.e., there is a linear dependency), then $\sigma(x, y)$ is 1 or -1 . If they are totally uncorrelated then $\sigma(x, y)$ is 0. Hence Eq. (4) represents the dissimilarity between x and y .

$$(1 - |\sigma(x, y)|) \quad (4)$$

Subsequently, according to the dissimilarity matrix a weighted complete graph G is formed. The value at row i and column j in the dissimilarity matrix S_m , represents the weight of the edge between node a_i and a_j . As each feature has some dissimilarity value with other features (present in dissimilarity symmetric matrix S_m), hence the graph G is a complete graph.

Now our problem is how to determine the optimal number of relevant and non-redundant attributes. So, we have to select an optimal subgraph g of G that contains the most relevant attributes and non-redundant. Therefore, we will apply the algorithm Multi-objective Binary PSO based Approach [6] as explained subsequently.

3.3 Subgraph Computation

The problem of finding the most relevant and non-redundant attributes is modeled by finding the optimal subgraph from a weighted undirected graph. The attributes contained by the extracted subgraph comprise the final selected relevant and non-redundant features.

Figure 3 is an example, that the left graph is the initial graph that contains 5 nodes and value of s_{ij} is the dissimilarity value between attributes and each attribute has a relevance value R_i .

The sub-graph on the right is the returned after the application of the optimization algorithm.

The problem of finding the subgraph is an NP-hard problem [1]. To solve our problem we will use multi-objective optimization algorithm graph based MbPSO [6], which is described by the Algorithm 2. In multi-objective optimization problem the objectives may estimate those different aspects of solutions which are conflicting in nature [6].

Algorithm 2. Graph based MOBPSo

```

1  Input: data matrix  $dt$ ,  $C$ =number of attributes,  $seuil=0.9$ .
2  Output: Archive  $A$ 
3  Begin
4       $[x_n, v_n, G_n, P_n]_{n=1}^N = \text{initialize}(dt)$                                  $\hookleftarrow$  Random locations and velocities
5       $g_n[V1, E1, VW1, EW1] = Sm(V1_n \times V1_n)$                              $\hookleftarrow$  subgraphs  $g_n$  for  $N$  particles are formed from dissimilarity matrix  $S_n$ 
6       $f1 = (1 - \frac{\sum_{i=1}^{|V1|} \sum_{j=1}^{|V1|} KW1_{ij}}{|V1|(|V1|-1)})^N$                          $\hookleftarrow$  average dissimilarity value for the  $N$  subgraphs
7       $f2 = (1 - \frac{\sum_{i=1}^{|V1|} VW1_i}{|V1|})^N$                                  $\hookleftarrow$  average snr value for the nodes contained by  $N$  subgraphs
8      IF  $fitness(x_n) \not\prec fitness(u), \forall u \in A$  THEN
9           $A := x_n$                                                      $\hookleftarrow$  Initialize archive  $A$  by first non-dominated  $x_n$ 
10     End IF
11     For  $n = 1..N$  do
12         For  $d = 1..C$  do
13              $v_{nd} := w \cdot v_{nd} + r1 \cdot (P_{nd} - x_{nd}) + r2 \cdot (G_{nd} - x_{nd})$ 
14              $x_{nd} := x_{nd} + v_{nd}$ 
15             IF  $x_{nd} \geq \text{seuil}$  THEN
16                  $x_{nd} := 1$                                              $\hookleftarrow$  discretize the cell value
17             End IF
18         End For
19     End For
20     For  $n = 1..N$  do
21          $g_n[V1, E1, VW1, EW1] = Sm(V1_n \times V1_n)$                      $\hookleftarrow$  new subgraph produced by the evaluated particles
22          $f1_n = \frac{\sum_{i=1}^{|V1|} \sum_{j=1}^{|V1|} KW1_{ij}}{|V1|(|V1|-1)}$                          $\hookleftarrow$  average dissimilarity value for the new subgraph
23          $f2_n = \frac{\sum_{i=1}^{|V1|} VW1_i}{|V1|}$                                  $\hookleftarrow$  average snr value the nodes contained by the new subgraph
24          $A := A \cup x_n$ 
25         For  $k = 1..N$  do
26             IF  $fitness(x_{nk}) \not\prec fitness(P_n)$  THEN
27                  $P_n := x_n$                                              $\hookleftarrow$  Update personal best
28             IF Non-dominated  $fitness$  THEN
29                  $P_n := \text{Random-choice}[x_n, P_n]$ 
30             End IF
31             End IF
32              $G_n := \text{random-select}(A)$ 
33         End For
34     End For
35     For
36          $A := x_n(S(fitness(x_{nk}) \not\prec fitness(P_n), \forall u \in A))$            $\hookleftarrow$  Non-dominated sorting is applied to the updated archive
37          $\text{CrowdingSort}(A)$                                              $\hookleftarrow$  crowding distance sorting for archive From step-6 to step-33 are repeated according to number of
38         iteration
39     End

```

The Algorithm 3 describes the different steps of our proposed new method cited above.

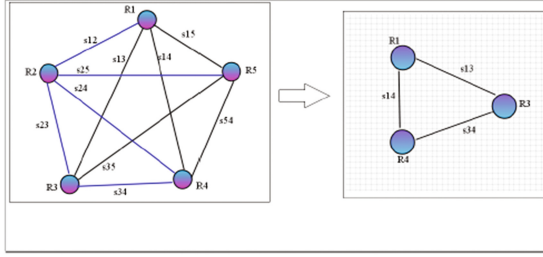


Fig. 3. Example of subgraph extraction

Algorithm 3. Reliable Attribute Slection based on Random forest(RASER)

```

1 Input:  $B = \{X_1, X_2, X_3, \dots, X_n\}$  : datasets
2  $Y = \{y_1, y_2, y_3, \dots, y_k\}$  : classes labels
3  $seuil=0.9$ 
4 Output: S:set of features
5 Begin
6    $R \leftarrow \text{Random Forest}(AT, B)$      $\triangleleft$  for each attribute , the degree of relevance is calculated
   using the Alogithme 1
7   For  $i = 1..n$  do
8     For  $j = 1..n$  do
9        $M_d(a_i, a_j) < -\sigma(a_i, a_j)$      $\triangleleft$  computes the dissimilarity matrix  $M_d$  using the
       proximity measure of the equation (4).
10    End For
11  End For
12   $G \leftarrow \text{generationGraph}(R, M_d)$      $\triangleleft$  generation Graph This is a function that generates a
  weighted graph.
13   $g \leftarrow \text{MbPSO}(B, X_i, Y, seuil, G)$ 
14 End

```

4 Experimental Results

In this section, we first describe the datasets, their preprocessing procedure and the performance metrics followed by the results of different algorithms.

In this article nine datasets are used which are publicly available from [1]. Table 1 summarizes these datasets. The first column reports the datasets name, the second column the number of features and the third column the number of instances. The final column reports the number of classes as well as the size of each class. For example, the WBC dataset is composed of 699 instances described each by 10 features. It is composed of two classes (458 Benign, 241 Malignant).

Performance is evaluated using sensitivity, specificity, accuracy and error-rate.

Table 1. Description of the used datasets

| Dataset name | # features | # instances | # classes |
|--------------|------------|-------------|---|
| WBC | 10 | 699 | 2 (458 Benign, 241 Malignant) |
| WPBC | 34 | 198 | 2 (151 non-r  current, 47 r  current) |
| WDBC | 32 | 569 | 2 (357 benign, 212 malignant) |
| Spambase | 57 | 4601 | 2 (spam, not spam) |
| Ionosphere | 34 | 351 | 2 (good, bad) |
| Madelon | 500 | 4400 | 2 |
| Dermatology | 33 | 366 | 6 (112 psoriasis, 61 seboreic dermatitis, 72 lichen planus, 49 pityriasis rosea, 52 cronic dermatitis, 20 pityriasis rubra pilaris) |

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (6)$$

$$\text{Errorrate} = \frac{FP + FN}{TN + TP + TN + FP + FN} \quad (7)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + TN + FP + FN} \quad (8)$$

Where

TP: True Positive = correctly identified

TN: True Negative = correctly rejected

FP: False Positive = incorrectly identified

FN: False Negative = incorrectly rejected

As this evaluate our method, it is necessary to use a datasets that were not used for learning: this is what is called the test datasets.

It contains also labeled examples to compare the predictions of a hypothesis with the actual value of the class. This test is generally achieved by reserving part of the initial supervised examples and that will not be used in the learning phase. Cross-validation is a method of estimating reliability of a model based on a sampling technique.

It is based on estimated performance from examples that were not used in the design of the model. The cross-validation algorithm k-fold cut the initial set of D examples in k blocks. The cross-validation algorithm is described in [1].

The performance of our proposed method RASER is compared with: MBEGA [3], mRMR(MID) [9], mRMR(MIQ) [9] methods. The datasets are arbitrarily divided into two groups: all together for learning and for testing. Each algorithm is executed for each training set and evaluated by the appropriate test portion.

The performance analysis is extended using 10-fold cross validation. The all algorithms are executed on the total sample versus datasets and the output features are validated using 10-fold cross-validation using Support Vector Machine (SVM).

4.1 Value of Relevance

To compute the features relevance value, we applied the random forest method that contains *ntree* trees. The Fig. 4 shows the influence on the choice of *ntree* value relevance. We can see in Fig. 4 that the number of trees (*ntree*) has practically no effect on the relevance.

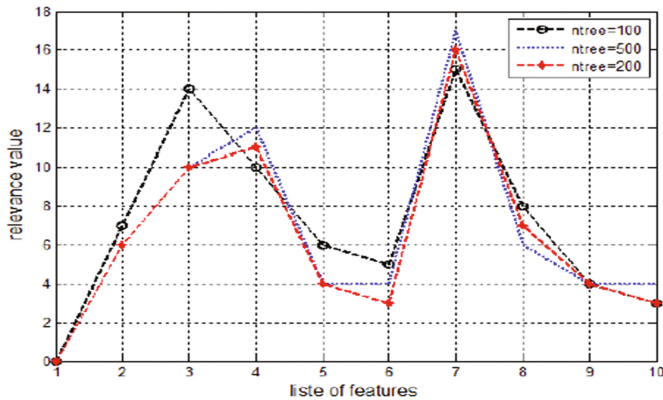


Fig. 4. Influence of *ntree* on value of relevance for WBC datasets

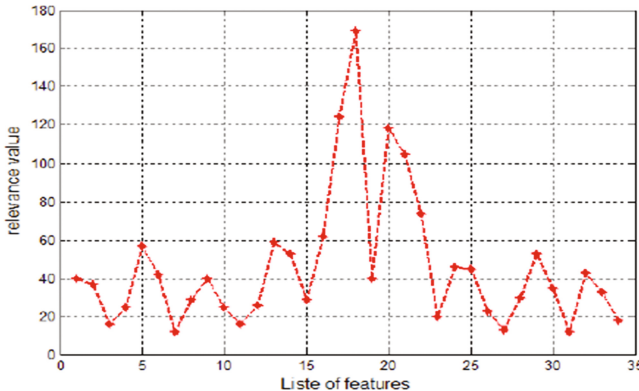


Fig. 5. Attributes relevance value of WPBC datasets

So for our work, we choose 200 as *ntree* value, that is, each built forest contains 200 trees. Figure 5 show the relevance of attribute values for the WPBC dataset.

4.2 Performance

In Table 2 we compare the classification results obtained with the original dataset that contains all the attributes and the dataset that contains the attributes selected by our method. We note that our algorithm, in general returns a classification rate higher than that returned by the classification of the original dataset.

Table 2. Comparisons of the classifications results

| Dataset | Methods | Specificity (%) | Sensitivity (%) | % accuracy | % error rate |
|-------------|-------------------------|-----------------|-----------------|-------------|--------------|
| WBC | RASER | 98 | 97.3 | 97.6 | 2.4 |
| | Original classification | 97 | 96.1 | 96.4 | 3.6 |
| BC | RASER | 88.1 | 96.1 | 93.4 | 6.5 |
| | Original classification | 55.3 | 78.1 | 76 | 24 |
| WPBC | RASER | 97.6 | 95.5 | 96.2 | 3.7 |
| | Original classification | 95.7 | 96.9 | 96.5 | 3.4 |
| WDBC | RASER | 95.5 | 98.3 | 97.2 | 2.7 |
| | Original classification | 93.3 | 96.5 | 95.3 | 4.7 |
| Spambase | RASER | 34 | 98 | 74 | 26 |
| | Original classification | 1 | 100 | 61 | 34 |
| Ionosphere | RASER | 96 | 97 | 96 | 3 |
| | Original classification | 52 | 66 | 64 | 36 |
| Dermatology | RASER | 99 | 100 | 95 | 5 |
| | Original classification | 100 | 98 | 89 | 10 |
| Madelon | RASER | 51 | 46 | 50 | 50 |
| | Original classification | 50 | 47 | 48 | 52 |

Table 3. Comparisons of the classifications results with other feature selection algorithms

| Dataset | Methods | Specificity (%) | Sensitivity (%) | % accuracy | % error rate |
|-------------|-----------|-----------------|-----------------|-------------|--------------|
| WBC | RASER | 98 | 97.3 | 97.6 | 2.4 |
| | MBEGA | 95.9 | 97.6 | 96.9 | 3.1 |
| | mRMR(MID) | 92 | 91.3 | 91.5 | 8.4 |
| | mRMR(MIQ) | 90.8 | 96.7 | 94.6 | 5.3 |
| BC | RASER | 88.1 | 96.1 | 93.4 | 6.5 |
| | MBEGA | 45.5 | 82.3 | 71.3 | 28.7 |
| | mRMR(MID) | 57.3 | 73.8 | 69.8 | 30.1 |
| | mRMR(MIQ) | 65.5 | 70.1 | 70.2 | 29.7 |
| WPBC | RASER | 97.6 | 95.5 | 96.2 | 3.7 |
| | MBEGA | 68.2 | 73.2 | 72.7 | 27.2 |
| | mRMR(MID) | 60.8 | 72.9 | 71.1 | 28.9 |
| | mRMR(MIQ) | 66.9 | 75.1 | 73.8 | 26.1 |
| WDBC | RASER | 95.5 | 98.3 | 97.2 | 2.7 |
| | MBEGA | 90.9 | 98.8 | 95.7 | 4.3 |
| | mRMR(MID) | 92.7 | 95.1 | 94.1 | 5.8 |
| | mRMR(MIQ) | 92.2 | 96.3 | 94.7 | 5.2 |
| Spambase | RASER | 34 | 98 | 74 | 26 |
| | MBEGA | 10 | 99 | 64 | 35 |
| | mRMR(MID) | 65 | 35 | 61 | 38 |
| | mRMR(MIQ) | 66 | 38 | 62 | 37 |
| Ionosphere | RASER | 96 | 97 | 96 | 3 |
| | MBEGA | 81 | 21 | 73 | 27 |
| | mRMR(MID) | 65 | 38 | 61 | 38 |
| | mRMR(MIQ) | 47 | 24 | 75 | 84 |
| Dermatology | RASER | 99 | 100 | 95 | 5 |
| | MBEGA | 100 | 97 | 90 | 10 |
| | mRMR(MID) | 97 | 96 | 81 | 18 |
| | mRMR(MIQ) | 98 | 96 | 93 | 6 |
| Madelon | RASER | 51 | 46 | 50 | 50 |
| | MBEGA | 38 | 57 | 54 | 46 |
| | mRMR(MID) | 5 | 54 | 0.53 | 47 |
| | mRMR(MIQ) | 49 | 51 | 50 | 50 |

The cross-validation scores of different algorithms are reported in Table 3. It is clear from this table for the most of datasets the sensitivity, specificity, accuracy, and errorrate values of our model are better than mRMR(MID) mRMR(MIQ) and MBEGA. The average Accuracy for our method is very much upper than other methods i.e. proposed method results are more non-redundant and relevant features than other comparative methods.

5 Conclusion

In this paper, we have proposed a novel feature selection algorithm based on random forest to select the most relevant and non-redundant one. In that purpose, we have used the random forest method to measure the relevance value attributes and the correlation coefficient to calculate the value of redundancy. The proposed method is performed in three basic steps three stages: the measure of relevance of attributes, modeling and optimization. Finally, we tested our approach for the selection of attributes on datasets and we have approved its effectiveness. These encouraging results should stimulate future research. Our immediate concern is to use another measures of dissimilarity criterion.

References

1. Samb, M.L., Camara, F., Ndiaye, S., Slimani, Y., Esseghir, M.A.: Approche de sélection d'attributs pour la classification basée sur l'algorithme rfe-svm
2. Chouaib, H.: Sélection de caractéristiques:méthodes et applications (2011). <http://www.math-info.univ-paris5.fr/~vincent/siten/Publications/theses/pdf/chouaib.pdf>
3. Zhu, Z., Ong, Y.-S., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recogn.* **40**, 3236–3248 (2007). <http://www.sciencedirect.com/science/article/pii/S0031320307000945>
4. John, G.H.: Enhancements to the data mining process. Doctoral dissertation, Ph.D. thesis of Stanford University (1997)
5. Kohavi, R., Pfleger, K., John, G.H.: Irrelevant features and the subset selection problem, pp. 121–129 (1994)
6. Mandal, M., Mukhopadhyay, A.: A graph-theoretic approach for identifying non-redundant and relevant gene markers from microarray data using multiobjective binary PSO. *PLoS ONE* **9**(3), e90949 (2014)
7. Koller, D., Sahami, M.: Toward Optimal Feature Selection. pp. 284–292. Stanford InfoLab, Stanford (1996)
8. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**(19), 2507–2517 (2007)
9. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of maxdependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005)
10. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **5**, 537–550 (1994)
11. You, W., Yang, Z., Ji, G.: PLS-based recursive feature elimination for high-dimensional small sample. *Knowl.-Based Syst.* **55**, 15–28 (2014)
12. Zhou, Q., Zhou, H., Zhou, Q., Yang, F., Luo, L.: Structure damage detection based on random forest recursive feature elimination. *Mech. Syst. Sig. Process.* **46**(1), 82–90 (2014)
13. Azhagusundari, B., Thanamani, A.S.: Feature selection based on information gain. *Int. J. Innov. Technol. Explor. Eng. (IJITEE)* ISSN 2278–3075 (2013)
14. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28 (2014)
15. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: *ICML*, vol. 3, pp. 856–863 (2003)

16. Ghattas, B., Ishak, A.B.: Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. *J. de la société française de statistique* **149**(3), 43–66 (2008)
17. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genomewide expression patterns. *Proc. Natl. Acad. Sci.* **95**(25), 14863–14868 (1998)
18. Crescenzi, P., Kann, V., Halldórsson, M.: A compendium of NP optimization problems (1995)
19. <https://archive.ics.uci.edu/ml/datasets.html>

Intelligent Systems Design and Applications
16th International Conference on Intelligent Systems
Design and Applications (ISDA 2016) held in Porto,
Portugal, December 16-18, 2016
Madureira, A.M.; Abraham, A.; Gamboa, D.; Novais, P.
(Eds.)
2017, XXI, 1086 p. 395 illus., Softcover
ISBN: 978-3-319-53479-4