RANDOM FOREST CLASSIFICATION AND ATTRIBUTE SELECTION – PROGRAM rfc3d

Contents

Overview

Random Forest (RF) is a supervised classification algorithm using multiple decision trees. Program **rfc3d** uses training data generated from facies interpretation or well log property and a number of seismic attributes as the input. The Output is predicted facies or class in 3D seismic volume. Attribute importance is a byproduct of RF algorithm, providing quantitative measure of how important or redundant each attribute is in the learning process. The program generates RF model with training data and cross-validate the model. The accuracy and feature importance are displayed.

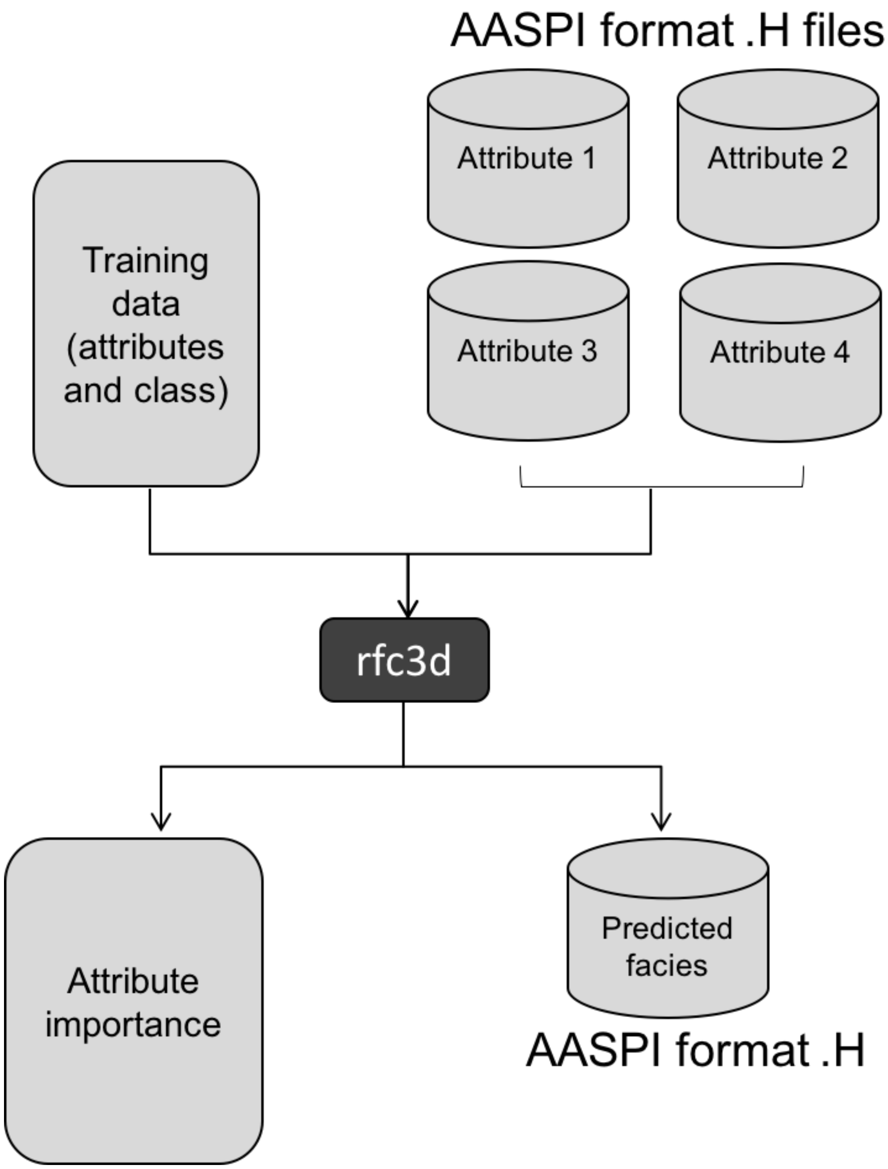


Figure 1. Workflow diagram of program rfc3d.

Random Forest and attribute importance

Single decision tree and random forests

Classification and regression tree (CART) is a machine leaning technique (Breiman et al, 1984).

In the CART algorithm, the best split is made using Gini impurity at each internal node in the tree given by

$$Gini\ i(\tau) = 1 - \sum_{\theta=1}^{k} p\,(\theta|\tau)^2\ ,$$

where k is the number of classes, and p(θ|τ) is the probability of class θ at node t. The probability is the fraction of observations that belong to class I at node t. The leaves are the final outcome or class. Gini impurity is a measurement of the likelihood of an incorrect classification of new instance, if it was randomly labeled according to the distribution of labels in the subset. Thus, the possible minimum value of Gini impurity is 0, when all observations belong to one class.

Prediction for N number of trees can be made by averaging the prediction of individual trees and is given by

$$i_{N_T}(\tau) = \frac{1}{N_T}\sum_T i(\tau)\ .$$

Feature importance

Selecting appropriate features is important in machine learning algorithms. Some features are more powerful for classification, and others may be redundant. Reduction of dimension based on feature selection can speed up the learning process, as well as improve prediction accuracy. To evaluate feature importance, Breiman et al. (2001, 2002) suggested Gini importance based on gini index as impurity function given by

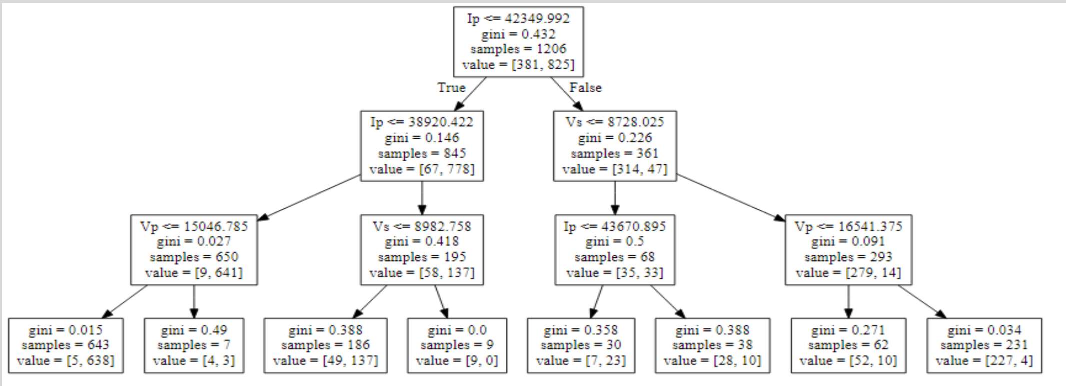
$$Gini\ Importance\ i_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, \mathrm{T}),$$

where, $\Delta i(\tau)$ is node purity gain which is denoted

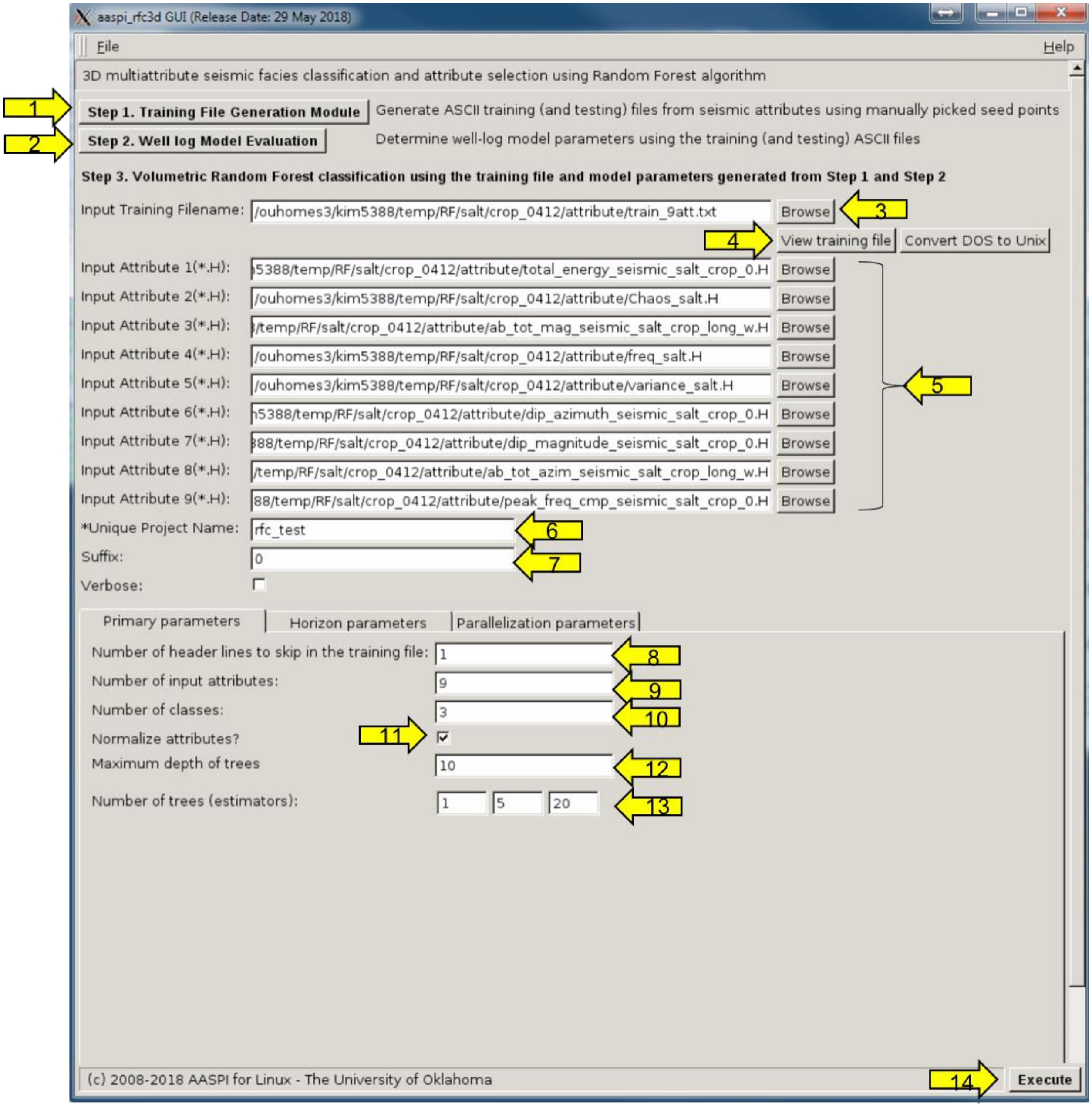
$$\Delta i(\tau) = i(\tau) - p_l i(\tau_l) - p_r i(\tau_r).$$

Decrease in Gini impurity, $\Delta i(\tau)$ results from splitting the samples to left and right sub-nodes.

Random Forest and attribute importance



Single decision tree example. A tree is composed of a root node, internal nodes and leaf nodes. The first split is made in the root node after an attribute of dataset is randomly chosen. The process repeats, splitting training set in internal nodes until it reaches leaf nodes when the node meets maximum depth or no more split can be made. The subset in leaf nodes returns class or label.



To start the program, click **rfc3d** in *Volumetric Classification* in the **aaspi_util** GUI, or type "**aaspi_rfc3d**" in terminal:

aaspi_rfc3d & rfc3d requires training file and multiple attributes as input. Training data can be generated with either (1) manually picked facies or (2) property interpreted from well log. The program **psvm3d** includes step-by-step instruction on generating training data.

→ Program psvm3d

Click (3) *Browse* and select text-format training file. To check the text file, click (4) *View training file*. (5) Choose AASPI format .H attribute files as inputs of facies classification. The output file is the same seismic volume as an input attribute format. Input (6) *unique project name* and (7) *Suffix*.

(8) Input the *Number of header lines you want to skip in the training file*. The default is 1. (9) Select *the number of input attributes*. The default value is updated when new attributes are added. Type (10) *number of classes* in training file. Toggle on (11) *Normalize attribute*, if scaling of each attribute is needed.

(12) *Maximum depth of tress* and (13) *Number of trees (estimators)* are hyper parameters which are related to random forest algorithm. Higher maximum depth makes more complicated tree model, but too higher depth cause overfitting. Random forest is an ensemble method using Bootstrap aggregation (Bagging). The method combines multiple decision trees, which enhance the accuracy of prediction. The higher number of estimators increases accuracy up to a certain point, but also increases the amount of computation. The program tests 3 cases of number of tree and adopt the best number of tree as a parameter for facies prediction. (14) Click *Execute* to start the program.

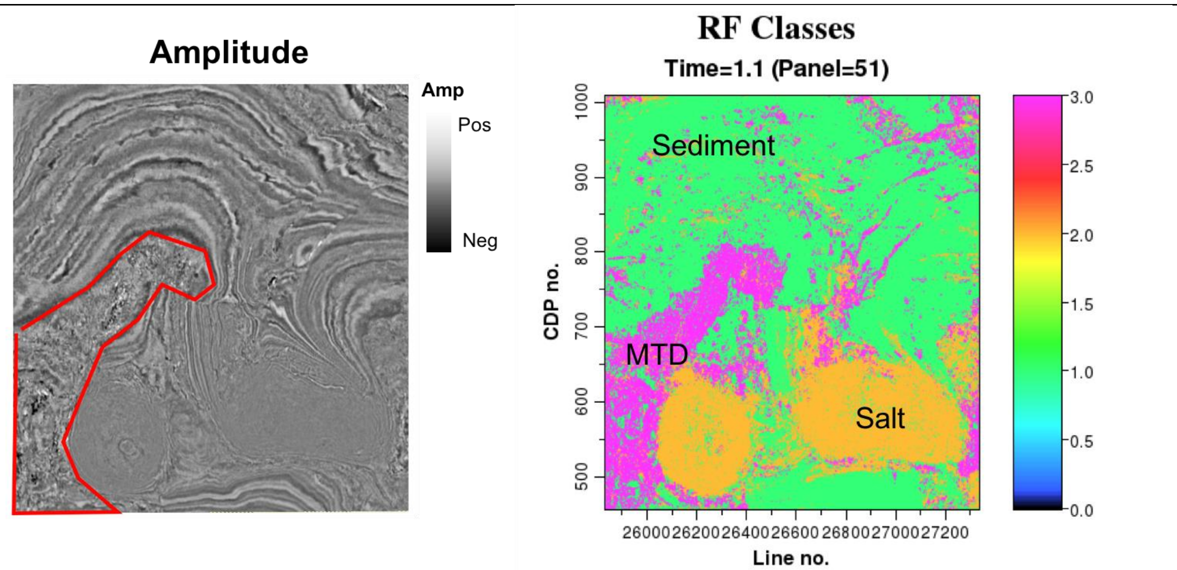
```
   ============================================================================
    Start 5-fold cross-validation of random forest classifer using training data
   ============================================================================
cross-validate with n number of trees:           1
fold number:           1 accuracy:  0.8405000
fold number:           2 accuracy:  0.8475000
fold number:           3 accuracy:  0.8260000
fold number:           4 accuracy:  0.8425000
fold number:           5 accuracy:  0.8275000
average accuracy with           1 decision trees:   0.8368000
cross-validate with n number of trees:           5
fold number:           1 accuracy:  0.8935000
fold number:           2 accuracy:  0.9000000
fold number:           3 accuracy:  0.8870000
fold number:           4 accuracy:  0.8935000
fold number:           5 accuracy:  0.8825000
average accuracy with           5 decision trees:   0.8913000
cross-validate with n number of trees:          20
fold number:           1 accuracy:  0.9040000
fold number:           2 accuracy:  0.9150000
fold number:           3 accuracy:  0.9090000
fold number:           4 accuracy:  0.9060000
fold number:           5 accuracy:  0.8975000
average accuracy with          20 decision trees:   0.9062999
best number of trees for classification=          20
   ============================================================================
                     feature importance of each attribute
   ============================================================================
    feature importance of attribute          1:          total energy broadband  0.1964
    feature importance of attribute          2:                           Chaos  0.1443
    feature importance of attribute          3:            Total aberrancy value  0.1011
    feature importance of attribute          4:         Instantaneous frequency  0.1189
    feature importance of attribute          5:                        Variance  0.1449
    feature importance of attribute          6:                     dip azimuth  0.1362
    feature importance of attribute          7:                   dip magnitude  0.0559
    feature importance of attribute          8:          Total aberrancy azimuth  0.0307
    feature importance of attribute          9:frequency at peak magnitude (cycles  0.0717
   ============================================================================
```

The program displays the result of 5-fold cross-validation of random forest model. After testing different number of trees, the tree number result in best accuracy is adopted for prediction. Using RF model generated from training data, attribute importance is evaluated and normalized importance is printed.

```
 ============================================================================
          read in attribute files and start RF classification
 ============================================================================
 0: first_line_out,current_line,last_line_out,ETA        1        1      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1        2      751      0.469 h
 0: first_line_out,current_line,last_line_out,ETA        1        3      751      0.351 h
 0: first_line_out,current_line,last_line_out,ETA        1        4      751      0.312 h
 0: first_line_out,current_line,last_line_out,ETA        1        5      751      0.292 h
 0: first_line_out,current_line,last_line_out,ETA        1        6      751      0.280 h
 0: first_line_out,current_line,last_line_out,ETA        1        7      751      0.272 h
 0: first_line_out,current_line,last_line_out,ETA        1        8      751      0.266 h
 0: first_line_out,current_line,last_line_out,ETA        1        9      751      0.261 h
 0: first_line_out,current_line,last_line_out,ETA        1       10      751      0.258 h
 0: first_line_out,current_line,last_line_out,ETA        1       11      751      0.255 h
 0: first_line_out,current_line,last_line_out,ETA        1       12      751      0.252 h
 0: first_line_out,current_line,last_line_out,ETA        1       13      751      0.250 h
 0: first_line_out,current_line,last_line_out,ETA        1       14      751      0.250 h
 0: first_line_out,current_line,last_line_out,ETA        1       15      751      0.249 h
 0: first_line_out,current_line,last_line_out,ETA        1       16      751      0.247 h
                                                    .
                                                    .
                                                    .
 0: first_line_out,current_line,last_line_out,ETA        1      744      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      745      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      746      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      747      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      748      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      749      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      750      751      0.000 h
 0: first_line_out,current_line,last_line_out,ETA        1      751      751      0.000 h
normal completetion. routine rfc3d.
```

When the model is generated, the facies prediction for the entire attribute volume starts and displays the inline currently processing and estimated time to complete. Predicted faces is output in format of AASPI .H file, which is named "**rfc3d_classification_<unique_project_name>.H**". The parameter information and print out from the program is stored in "**aaspi_rfc3d_<unique_project_name>.out**".

Example



The representative time slice through seismic amplitude (left) and predicted class (right). The prediction was made with RF algorithm using 9 attribute volumes. The area bounded with the red line in the left image has chaotic and discontinuous seismic expression, which is interpreted as mass transport deposits.

References

Breiman, L., J., Friedman, C. J. Stone, and R. A. Olshen, 1984, Classification and regression trees: CRC press.

Breiman, L., 2001b. Random forests. Mach. Learning **45**, 5–32.

Liaw, A. and M., Wiener, 2002. Classification and regression by random Forest: R News **2** - 3, 18–22.

Sandri, M. and P., Zuccolotto, 2008. A bias correction algorithm for the Gini variable importance measure in classification trees. Journal of Computational and Graphical Statistics, **17**-3, 611-628.

Menze, B. H., B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics, **10**-1, 213.