QuickStart Manual for Decision Forest

Decision Forest (DF) is a consensus modeling technique that combines multiple Decision Tree models in a manner that results in more accurate predictions. Since combining several identical trees produces no gain, the rationale behind DF is use of individual trees that are different (i.e., heterogeneous) in representing the association of independent variables (gene expression in DNA microarray, m/z peaks in SELDI-TOF data and structural descriptors in SAR modeling) and dependent variable (class categories) yet comparable in their prediction accuracy. The heterogeneity requirement assures that each tree uniquely contributes to the combined prediction, while the quality comparability requirement assures that each tree makes a similar contribution to the combined prediction. Since a certain degree of noise is always present in biological data, optimizing a tree inherently risks over fitting the noise. DF tries to minimize overfitting by maximizing the difference among individual trees in order to cancel some random noise in individual trees. The maximum difference was achieved by constructing each individual tree using a distinct set of dependent variables.

This version of Decision Forest is only suitable for two-class fitting model development and prediction.

1. Develop a forest fitting model

1. Start with a dataset that is the tab delimited text file with column names on the first row. We recommend (but don't have to) that the first one or two columns contain sample (or chemical) IDs, followed by the dependent variable and independent variables (see the sample file).

Figure 1

2. Select Modeling >Build Forest (Figure 1).

3. Browse and select the dataset, click "Select data file".

4. In the popup window, select the dependent and independent variables (Figure 2). Then click "OK". The forest model is developed and automatically saved to the directory where the dataset is located. The model file has the same name as the dataset with the extension .mdl.

Figure 2

5. The forest result is displayed in the left panel of the window as a folder (Figure 3). The fold can be expanded by double clicking to show the results of individual tree as well as the summary information.
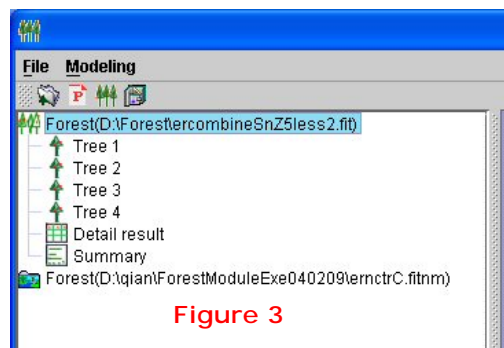

Figure 3

2. Prediction

The forest model can be used for prediction of unknown samples (or chemicals). The predicted dataset should also be the tab delimited text file with the independent variables.

1. Select Modeling >Prediction.

2. In the popped up dialog (Figure 4), either browse for a saved forest model or choose an opened forest model, and then browse and select the dataset for prediction. Then, click "Next".

3. In the next popped up window (Figure 5), select the independent variables from the dataset. Click "OK".

4. The prediction result is shown as a folder in the left panel of the main window.
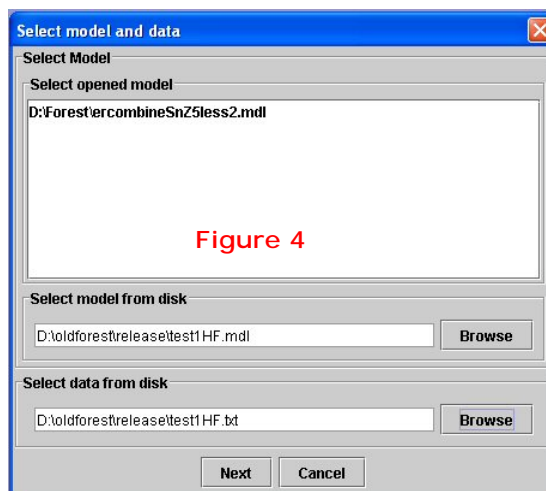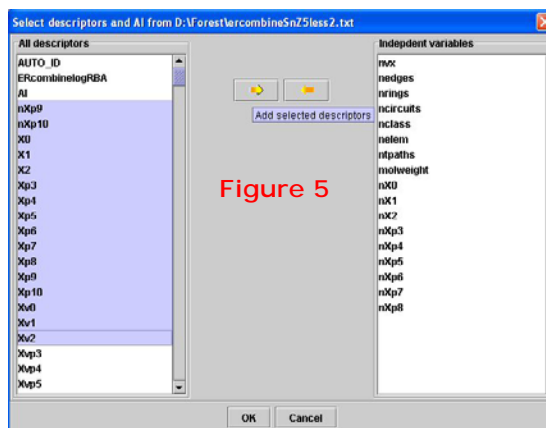

Figure 4

Figure 5

3. Other features

View fitting model and prediction results (Figure 6):

1. Select File>Load Forest Fitting result (*.fit) to view the fitting results.

2. Select File>Load Prediction Result (*.pre) to view the prediction results.

Change the parameters for the tree development:
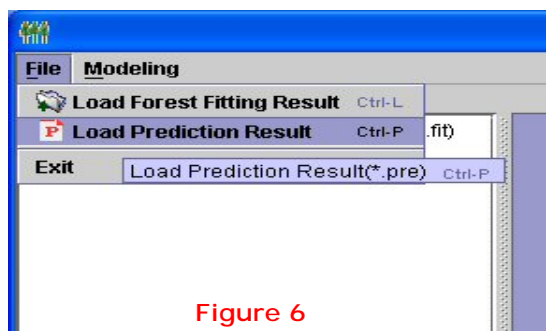
1. Select Modeling>Parameters.

Figure 6