User Guide for FBPtree

1 Introduction

FBPtree is a classification tree software implementing the Frontier-Based Tree-Pruning (FBP) algorithm. The algorithm is described in Huo et al. (2004). The software is coded in Matlab and consists of the following components:

- **Data Preprocessing.**
  Preprocess.m

- **Tree Growing.**
  Treegrow.m, bestsplit.m, impurity.m, count.m

- **Tree Pruning.**
  Treepruning.m, tmerge.m

- **\( f(\lambda) \) identification, where \( f(\lambda) \) is the minimum complexity-penalized loss function for given penalizing parameter \( \lambda. \)**
  FindLowerBound.m

- **Best tree identification for any given \( \lambda. \)**
  Treesize.m, Bestree.m

- **Testing.**
  Test.m, Prediction.m

- **Application in Cross Validation.**
  FBPtree.m, FBPtest.m, CVtree.m, CombPLCurves.m, ExtractError.m

2 Usage

To use FBPtree, type the following in the command window of Matlab.

\[
[\text{bestree} \_\text{size}, \text{bestree} \_\text{mat}, \text{best} \_\text{split}, \text{error}, \text{pred}] = \text{FBPtree}(\text{train, data} \_\text{type, test, rule})
\]
2.1 Input

The inputs to FBPtr tree include:

- **train** - the training sample, a $n \times m$ matrix, where $n$ is the training sample size and $m$ is the number of variables.

- **data_type** - the types of each variable, a $1 \times m$ character vector consists of as follows:
  - d for a dependent variable. One and only one variable must have the d indicator.
  - n for a numerical predictive variable.
  - c for a categorical predictive variable.
  - e if a variable is excluded from the analysis.

For instance, the data matrix for the Iris Plant example has five columns. The first four columns contain the four numerical predictive variables, sepal length, sepal width, petal length, and petal width, respectively, and the last column gives the classes of plants. Therefore, for the Iris Plant example, data_type = 'nnnnd'.

- **test** (optional) - the testing sample, a $n' \times m$ matrix, where $n'$ is the size of the testing sample, and $m$ is the number of variables.

- **rule** (optional) - the splitting rule (1 for Misclassification Rate; 2 for Gini Index; 3 for Entropy). The default splitting rule is 1, i.e., Misclassification Rate.

2.2 Output

The outputs from FBPtr tree include:

- **bestree_size** - the size of the best tree, i.e., the total number of terminal nodes of the best tree.

- **bestree_mat** - a matrix recording the structure of the best tree. Each row of bestree_mat contains: node number, left child, right child, number of terminal nodes below, number of misclassifications, and class.

- **best_split** - a cell array recording the splitting information of the best tree. Each cell of best_split contains: node number, splitting variable, and splitting position.

- **error** - misclassification rate on the testing sample. If the testing sample is not available, misclassification rate on the training sample is given.
• **pred** - predicted classes for the testing sample. If the testing sample is not available, predicted classes for the training sample are given.

### 2.3 Illustrations

`work04.m` illustrates how to use the FBPtree software. The Iris Plant data is used for the illustration. In addition, `work01.m`, `work02.m`, and `work03.m` are the `.m` files for generating Figures 5, 6, and 7, respectively in Huo et al. (2004).

### References