A short `fscaret` package introduction with examples

Jakub Szlek (j.szlek@uj.edu.pl)

March 27, 2014

1 Installation

As it is in case of `caret`, the `fscaret` uses large number of R packages but it loads them when needed. To fully take advantage of the package it is recommended to install it both with dependent and suggested packages. Install `fscaret` with the command

```
> install.packages("fscaret", dependencies = c("Depends", "Suggests"))
```

from R console.

2 Overview

In general `fscaret` is a wrapper module. It uses the engine of `caret` to build models and to get the variable ranking from them. When models are build package tries to draw variable importance from them also getting the generalization error (RMSE and MSE) which is used during the scaling process. Finally the output is produced. It contains the data frame of variable importance, generalization error for all build models and preprocessed data set if the `preprocessData = TRUE` when calling the main `fscaret()` function.

In summary the whole feature ranking process can be divided into:

1. User provides input data sets and a few settings
2. Models are build
3. Variable rankings are draw out of the models
4. Generalization error is calculated for each model
5. Variable rankings are scaled according to generalization error

6. The results are gathered in tables

3 Input

3.1 Data set format

Be advised that fscaret assumes that data sets are in MISO format (multiple input single output). The example of such (with header) is:

<table>
<thead>
<tr>
<th>Input_no1</th>
<th>Input_no2</th>
<th>Input_no3</th>
<th>...</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.1</td>
<td>32.06</td>
<td>...</td>
<td>1.02</td>
</tr>
<tr>
<td>5</td>
<td>1.21</td>
<td>2.06</td>
<td>...</td>
<td>7.2</td>
</tr>
</tbody>
</table>

For more information on reading files in R, please write `?read.csv` in R console.

3.2 An example

There are plenty of methods to introduce data sets into R. The best way is to read file (presumably csv with tab as column separator) as follows:

1. Select file name

   ```
   > basename_file <- "My_database"
   > file_name <- paste(basename_file,".csv",sep="")
   ```

2. Read data file into matrix

   ```
   > matrixTrain <- read.csv(file_name,header=TRUE,sep="\t",
   + strip.white = TRUE, na.strings = c("NA"",""))
   ```

3. Put loaded matrix into data.frame

   ```
   > matrixTrain <- as.data.frame(matrixTrain)
   ```

   Be advised to use `header=TRUE` when you have data set with column names as first row and `header=FALSE` when there are no column names. The last step is obligatory to introduce data into fscaret functions as it checks if the data presented is in data.frame format.
4 Function \texttt{fscaret()}

4.1 Settings

All the settings are documented in Reference manual of \texttt{fscaret} \url{http://cran.r-project.org/web/packages/fscaret/fscaret.pdf}. Here we will concentrate only on a few valuable ones.

- \texttt{installReqPckg} The default setting is \texttt{FALSE}, but if set to \texttt{TRUE} prior to calculations it installs all packages from the sections ‘Depends’ and ‘Suggests’ of DESCRIPTION. Please be advised to be logged as root (admin) if you want to install packages for all users.

- \texttt{preprocessData} The default setting is \texttt{FALSE}, but if set to \texttt{TRUE} prior to calculations it performs the data preprocessing, which in short is realized in two steps:

  1. Check for near zero variance predictors and flag as near zero if:
     - the percentage of unique values is less than 20
     - the ratio of the most frequent to the second most frequent value is greater than 20,
  2. Check for susceptibility to multicollinearity
     - Calculate correlation matrix
     - Find variables with correlation 0.9 or more and delete them

- \texttt{regPred} Default option is \texttt{TRUE} and so the regression models are applied

- \texttt{myTimeLimit} Time limit in seconds for single model development, be advised that some models need as time to be build, if the option is omitted, the standard 24-hours time limit is applied. This function is off on non-Unix like systems.

- \texttt{Used.funcRegPred} Vector of regression models to be used, for all available models please enter \texttt{Used.funcRegPred="all"}, the listing of functions is:

```r
> library(fscaret)
> data(funcRegPred)
> funcRegPred
```

```
[1] "glm"      "glmStepAIC" "gam"      "gamLoess"
[5] "gamSpline" "rpart"     "rpart2"   "ctree"
[9] "ctree2"   "evtree"    "obliqueTree" "gbm"
[13] "blackboost" "bstTree" "glmboost" "gamboost"
[17] "bstSm"    "rf"        "parRF" "RRF"
[21] "cforest"  "Boruta"   "RRFglobal" "bagEarth"
[25] "treebag"  "bag"      "logicBag" "gcvEarth"
[29] "nodeHarvest" "partDSA" "earth"    "mlp"
[33] "logreg"   "glmnet"   "nnet"     "mlpWeightDecay"
[37] "mlpWeightDecay" " pcaNNet" " avNNet" " rbf"
[41] "pls"      "kernelpls" "simpls"   "widekernelpls"
```
A short fscaret package introduction with examples

- **no.cores** The default setting is **NULL** as to maximize the CPU utilization and to use all available cores. This option is off for Windows OS.

- **missData** This option handles the missing data. Possible values are:
  - missData="delRow" - for deletion of observations (rows) with missing values,
  - missData="delCol" - for deletion of attributes (columns) with missing values,
  - missData="meanCol" - for imputing mean to missing values,
  - missData=NULL - no action is taken.

- **supress.output** Default option is **FALSE**, but it is sometimes justifed to supress the output of intermediate functions and focus on ranking predictions.

### 4.2 An example

A simple example utilizing the data provided in the **fscaret**:

```r
> library(fscaret)
> data(dataset.train)
> data(dataset.test)
> trainDF <- dataset.train
> testDF <- dataset.test
> myFS<-fscaret(trainDF, testDF, myTimeLimit = 15, preprocessData=TRUE,
>                Used.funcRegPred=c("pcr","pls"), with.labels=TRUE,
>                supress.output=TRUE, no.cores=1)
> myRES_tab <- myFS$VarImp$matrixVarImp.MSE[1:10,]
> myRES_tab <- subset(myRES_tab, select=c("pcr","pls","SUM%","ImpGrad","Input_no"))
> myRES_PPlabels <- myFS$PPlabels
> myRES_PPlabels
```
5 Output

As it was stated previously there are three lists of outputs.

1. Feature ranking and generalization errors for models:

   > # Print out the Variable importance results for MSE scaling
   > print(myRES_tab)

<table>
<thead>
<tr>
<th>pcr</th>
<th>pls</th>
<th>SUM%</th>
<th>ImpGrad</th>
<th>Input_no</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>5.862841e+01</td>
<td>5.227714e+01</td>
<td>1.000000e+02</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1.567799e+01</td>
<td>2.741963e+01</td>
<td>3.865975e+01</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>1.916511e+01</td>
<td>1.995465e+01</td>
<td>3.527303e+01</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>2.519981e-01</td>
<td>3.161112e-01</td>
<td>5.122461e-01</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>1.872058e-02</td>
<td>3.079973e-02</td>
<td>4.465089e-02</td>
</tr>
<tr>
<td>13</td>
<td>22</td>
<td>2.880832e-04</td>
<td>1.324904e-03</td>
<td>1.454379e-03</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>5.880416e-04</td>
<td>2.781880e-04</td>
<td>7.810516e-04</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>7.190168e-05</td>
<td>6.894892e-05</td>
<td>1.270005e-04</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>1.570926e-06</td>
<td>2.697715e-06</td>
<td>3.848898e-06</td>
</tr>
<tr>
<td>21</td>
<td>13</td>
<td>1.081909e-06</td>
<td>1.078743e-06</td>
<td>1.948191e-06</td>
</tr>
</tbody>
</table>

2. Raw generalization errors for each model

   > # Print out the generalization error for models
   > print(myRES_rawMSE)

   pcr   | pls
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>716.6597</td>
</tr>
</tbody>
</table>

   > # Print out the reduced number of inputs after preprocessing
   > print(myRES_PPlabels)

<table>
<thead>
<tr>
<th>Orig Input No</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Balaban.index</td>
</tr>
<tr>
<td>2</td>
<td>Dreiding.energy</td>
</tr>
<tr>
<td>3</td>
<td>Fused.aromatic.ring.count</td>
</tr>
<tr>
<td>4</td>
<td>Hyper.wiener.index</td>
</tr>
<tr>
<td>5</td>
<td>Szeged.index</td>
</tr>
<tr>
<td>6</td>
<td>Ring.count.of.atom</td>
</tr>
<tr>
<td>7</td>
<td>pI</td>
</tr>
<tr>
<td>8</td>
<td>Quaternary_structure</td>
</tr>
<tr>
<td>9</td>
<td>PLGA_Mw</td>
</tr>
<tr>
<td>10</td>
<td>La_to_Gly</td>
</tr>
<tr>
<td>11</td>
<td>PVA_conc_inner_phase</td>
</tr>
<tr>
<td>12</td>
<td>PVA_conc_outer_phase</td>
</tr>
<tr>
<td>13</td>
<td>PVA_Mw</td>
</tr>
<tr>
<td>14</td>
<td>Inner_phase_volume</td>
</tr>
</tbody>
</table>
A short \texttt{fscaret} package introduction with examples

Figure 1: A sum of feature ranking of models trained and tested on \texttt{dataset.train}, two models were used "pcr","pls".

As one can see in the example there were only two models used "pcr","pls", to use all available models please set option \texttt{Used.funcRegPred="all"}. The results can be presented on a bar plot (see Figure 1). Then the arbitrary feature reduction can be applied.

6 Acknowledgments

This work was funded by Poland-Singapore bilateral cooperation project no 2/3/POL-SIN/2012.
A short fscaret package introduction with examples

7 References

