

A short **fscaret** package introduction with examples

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

October 6, 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.

Be advised! Running above code would install all possible packages (in some cases more than 240!), but it is necessary to fully benefit from **fscaret**. If you wish to use only specific algorithms, check which parameter from **funcRegPred** corresponds to which package.

In second case install **fscaret** with the command

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

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

The `fscaret` version 0.8.6.2 does not support methods for variable importance ranking harvesting! Next update (v0.9) will have this functionality.

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:

Input_no1	Input_no2	Input_no3	...	Output
2	5.1	32.06	...	1.02
5	1.21	2.06	...	7.2

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 `fscaret()`

4.1 Settings

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

- `installReqPckg` The default setting is `FALSE`, but if set to `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.
- `preprocessData` The default setting is `FALSE`, but if set to `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
- `regPred` Default option is `TRUE` and so the regression models are applied
- `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.
- `Used.funcRegPred` Vector of regression models to be used, for all available models please enter `Used.funcRegPred="all"`, the listing of functions is:

```

> library(fscaret)
> data(funcRegPred)
> funcRegPred

[1] "ANFIS"           "avNNet"          "bag"
[4] "bagEarth"        "bayesglm"        "bdk"
[7] "blackboost"      "Boruta"          "bstLs"
[10] "bstSm"           "bstTree"         "cforest"
[13] "ctree"           "ctree2"          "cubist"
[16] "DENFIS"          "dnn"             "earth"
[19] "elm"             "enet"            "evtree"
[22] "extraTrees"      "FIR.DM"          "foba"
[25] "FS.HGD"          "gam"             "gamboost"
[28] "gamLoess"        "gamSpline"       "gaussprLinear"
[31] "gaussprPoly"     "gaussprRadial"   "gbm"
[34] "gcvEarth"        "GFS.FR.MOGAL"    "GFS.LT.RS"
[37] "GFS.Thrift"      "glm"             "glmboost"
[40] "glmnet"          "glmStepAIC"      "HYFIS"
[43] "icr"             "kernelpls"       "kknn"
[46] "knn"             "krlsPoly"        "krlsRadial"
[49] "lars"            "lars2"           "lasso"
[52] "leapBackward"    "leapForward"     "leapSeq"
[55] "lm"             "lmStepAIC"       "logicBag"
[58] "logreg"          "M5"              "M5Rules"
[61] "mlp"            "mlpWeightDecay"  "neuralnet"
[64] "nnet"           "nodeHarvest"     "parRF"
[67] "partDSA"        "pcaNNet"         "pcr"
[70] "penalized"      "pls"             "plsRglm"
[73] "ppr"            "qrf"             "qrnn"
[76] "rbfDDA"         "relaxo"          "rf"
[79] "ridge"          "rknn"            "rknnBel"
[82] "rlm"            "rpart"           "rpart2"
[85] "RRF"           "RRFglobal"       "rvmlLinear"
[88] "rvmPoly"        "rvmRadial"       "SBC"
[91] "simpls"         "splS"            "superpc"
[94] "svmBoundrangeString" "svmExpoString"  "svmLinear"
[97] "svmPoly"        "svmRadial"       "svmRadialCost"
[100] "svmSpectrumString" "treebag"        "widekernelpls"
[103] "WM"            "xyf"

```

- `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 justified to suppress the output of intermediate functions and focus on ranking predictions.

4.2 An example

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

```
> 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_rawMSE <- myFS$VarImp$rawMSE
> myRES_PPlabels <- myFS$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)
```

	pcr	pls	SUM%	ImpGrad	Input_no
4	7.434019e+01	5.061400e+01	1.000000e+02	0.00000	4
5	1.829938e+01	2.640665e+01	3.577794e+01	64.22206	5
22	1.492109e+00	2.249200e+01	1.919432e+01	46.35152	22
23	2.052572e-01	4.426772e-01	5.185376e-01	97.29848	23
2	2.312969e-02	4.252871e-02	5.254598e-02	89.86651	2
13	8.167511e-05	1.651340e-03	1.386920e-03	97.36056	13
9	4.886944e-04	3.839307e-04	6.983560e-04	49.64700	9
1	7.732202e-05	9.634143e-05	1.389817e-04	80.09873	1
17	1.275905e-06	3.499222e-06	3.821502e-06	97.25036	17
21	9.553480e-07	1.459590e-06	1.932659e-06	49.42673	21

2. Raw generalization errors for each model

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

      pcr      pls
1 759.6954 716.854

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

preprocessRes$labelsDF[, ncol(preprocessRes$labelsDF)]
1                      Balaban.index
2                      Dreiding.energy
3      Fused.aromatic.ring.count
4                      Hyper.wiener.index
5                      Szeged.index
6      Ring.count.of.atom
7                      pI
8      Quaternary_structure
9                      PLGA_Mw
10                     La_to_Gly
11      PVA_conc_inner_phase
12      PVA_conc_outer_phase
13                      PVA_Mw
14      Inner_phase_volume
15                     Encaps_rate
16                     PLGA_conc
17      PLGA_to_Placticizer
18                     diss_pH
19                     diss_add
20                     Prod_method
21      Asymmetric.atom.count.1
22      Hyper.wiener.index.1
23                     Szeged.index.1
24                     count
25                     pH_14_logd
26                     bpKa2
27      Cyclomatic.number.2
```

As one can see in the example there were only two models used "pcr","pls", to use all available models please set option `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.

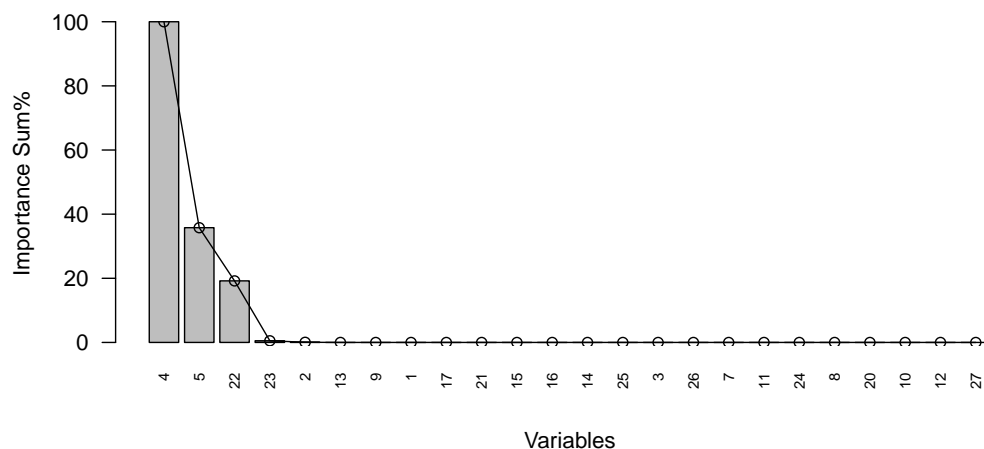


Figure 1: A sum of feature ranking of models trained and tested on `dataset.train`, two models were used "pcr","pls".

7 References

1. Szlek J, Paclawski A, Lau R, Jachowicz R, Mendyk A. Heuristic modeling of macromolecule release from PLGA microspheres. *International Journal of Nanomedicine*. 2013;8(1); 4601 - 4611. [link to webpage](#)
2. Szlek, J., Paclawski, A., Lau, R., Jachowicz, R., Mendyk, A. Heuristic modeling of macromolecules release from PLGA microspheres. *Conference proceedings*. Gdansk, May 24-25, 2013.[Abstract book](#)
3. Szlek J., Paclawski A., Lau R., Wisniowska B., Polak S., Jachowicz R., Mendyk A. The `fscaret`: automated feature ranking package. Under review.