IRootLab Tutorials

Feature histograms

Julio Trevisan – juliotrevisan@gmail.com

31th/July/2013


This document is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](http://creativecommons.org/licenses/by-nc-sa/3.0/).

Introduction 1

Creating objects 1

Saving the procedure so far for another occasion (optional) 7

Loading and preparing dataset 7

Applying Repeated feature selection (RFS) 10

Generating the histogram 11

# Introduction

Feature histograms are obtained through repeating feature selection several times (*e.g.*, 100) and counting how much each feature (*i.e.*, wavenumber) was selected. The procedure is applied to 2-class dataset to discover which features are most important for distinguishing between a condition and a reference class (*e.g.*, “treated” *vs.* “control”).

The feature selection method employed is called Forward feature selection (FFS). It uses a classifier to guide the addition of new features [1]. Starting with an empty feature set, the feature that gives best classification is chosen. Then, it chooses the feature that gives best classification together with the already chosen one, and so on. Training is done using a random 90% portion of the data and testing is done using the remaining 10% of the data.

# Creating objects

1. Start MATLAB and IRootLab as indicated in IRootLab manual (<http://irootlab.googlecode.com>).
2. In MATLAB command prompt, type “objtool”.
3. Click on “Block” in left panel.
4. Click on “New…” in middle panel

 

1. Locate and double-click on “Gaussian fit”



1. Click on “OK”.



1. Click on “Feature Subset Graded” in left panel.
2. Click on “New…” in middle panel



1. Locate and double-click on “using Classifier”



1. Select “clssr\_d01” in the “Classifier” drop-down box.



1. Click on “Sub-dataset Generation Specs” in left panel.
2. Click on “New…” in middle panel.



1. Double-click on “Random Sub-sampling”



1. Enter “100” in the “Number of repetitions” box.
2. Click on “OK”.



1. Enter any number on the “Random seed” box, *e.g.*, “121212”. This guarantees the exact repeatability of the results (if ever needed).



1. Click on “Block” in left panel.
2. Click on “New…” in middle panel.



1. Locate and double-click on “Forward”.



1. Enter “5” in the “Number of variables” box.
2. Select “fsg\_clssr01” in the “Feature Subset Grader” drop-down menu.



1. Click on “New…” again.



1. Locate and double-click on “Feature Selection Repeater”



1. Select “as\_fsel\_forward01” in the “Feature Selection” box.
2. Select “sgs\_randsub01” in the “SGS” box.



## Saving the procedure so far for another occasion (optional)

All the objects created up to step 25 are recorded in an automatically generated MATLAB script. If you run this script, it will automatically reproduce steps 1 to 25. This can be useful, *e.g.*, next time you start MATLAB.

1. In MATLAB command window, type in “edit\_ircode”. MATLAB editor will open with the corresponding script.
2. Save the file with another name, *e.g.*, “feahist\_objects” in a location of your preference.
3. Next time you start MATLAB, change your Current folder to the location of your preference above and type in the name of the file that you created in MATLAB command window, then open objtool and you will see that all objects have already been created. So, this time you will be able to start at step 29 below.

# Loading and preparing dataset

This tutorial uses Ketan’s Brain dataset supplied together with IRootLab.

1. In MATLAB command window, type “browse\_demos”



1. Locate and click on the “LOAD\_DATA\_KETAN\_BRAIN\_ATR” link.



1. Go back to objtool.
2. Click on “Refresh”.



1. Click on “ds01” in middle panel
2. Locate and double-click on “One-versus-reference” (sub-item of “Split”, not “Ensemble”)



Note that this dataset has three classes (in this order):

* Normal
* Glioblastoma
* Astrocytoma

“Index of reference class” = “1” refers to the first class (“Normal”), which will be considered as the reference class

1. Click on “OK”.



This will create two datasets with the following classes:

1. “ds01\_ovr01\_01” with classes “Normal” and “Glioblastoma”
2. “ds01\_ovr01\_02” with classes “Normal” and “Astrocytoma”

**Hint:** it is possible to check the “title” property of each dataset (Click on “Object properties” on the top right corner of objtool). The “title” property will be something like “Glioblastoma vs. Normal”.

# Applying Repeated feature selection (RFS)

The rule-of-thumb before a classification operation is to standardize the dataset to improve numerical stability of calculations.

1. Click on “Dataset” in left panel.
2. Click on “ds01\_ovr01\_01” in middle panel.
3. Locate and double-click on “Standardization” in right panel.



1. Click on “ds01\_ovr01\_01” in middle panel.
2. Click on “Existing blocks” (top right).
3. Double-click on “fselrepeater01” in right panel. **Warning:** this will start RFS, which is a time-consuming operation (took 536 seconds on a Core I7 @ 2GHz computer running MATLAB on Linux).



# Generating the histogram

1. Click on “Log” in left panel.
2. Click on “log\_fselrepeater\_fselrepeater01” in middle panel.
3. Locate and double-click on “Feature subsets processor” in right panel.



There are many parameters affecting the way the multiple feature sub-sets from the previous step are combined into a histogram. The most straightforward way is to just count how many times each feature was selected. This is done in this tutorial. However, other options can be explored in the box below by changing parameters and clicking on “Preview”.

1. Click on “OK”



1. Click on “log\_hist\_subsetsprocessor01” in middle panel.
2. Double-click on “Stacked Histograms” in right panel.



1. Click on “OK”.

**Note:** a Peak detector could be passed here (needs to be created first) to mark the peaks of the histogram.

**Note2:** The “Colors” box has a syntax of its own, explained at <http://bioph.lancs.ac.uk/irootdoc/html/colors2map_8m.html> . The simplest edit to this property is to replace “[.85, 0, 0]” (darkish red) with other combination of Red, Blue and Green proportions.

This results in the following figure. The figure is a histogram of the most important features for the “Glioblastoma vs. Normal” case.

# Validation

The classification rates have to be good in order to the interpretation of the histogram to be valid. This section verifies how the average classification rate improves as new features are added during feature selection. Because feature selection is repeated 100 times, the result here will be an average curve with standard deviations.

1. Click on “Log” in left panel.
2. Click on “log\_fselrepeater\_fselrepeater01” in middle panel.
3. Double-click on “extract\_dataset\_nfxgrade” in right panel.



1. Now click on “Dataset” in left panel.
2. Click on “irdata\_fselrepeater01” in middle panel.
3. Double-click on “Class means with standard deviation” in right panel.



This will result in the following figure:



This shows that the classification rate achieves 100% (don’t mind the y-axis label; it is classification rates (in %) being shown here) once 3 features are added. Because this is a high value (actually ideal for this dataset), it legitimizes the interpretation of the histogram for biomarkers identification.

1. Repeat steps 37-54 for the “ds01\_ovr01\_02” dataset (instead of “ds01\_ovr01\_01”). This covers the “Astrocytoma vs. Normal” case.