Feature Selection (FS) workflow report

May 18, 2017

Introduction

The report summarizing the Feature Selection pipeline results.

Feature Selection workflow

Naive Random Forest (RF) implementation for variable importance evaluation.

Dataset

Expression data from normal and prostate tumor tissues (GSE6919_GPL92).

Summary stats from training phase

```
## Information not available for this FS workflow implementation
```

Summary stats from testing phase

Table 1: Classification metrics from twenty class-balanced and randomized runs.

run	Variables	Accuracy	Kappa	AccuracyPValue
1	4910	0.6727	0.5055	3.979e-05
2	4833	0.6	0.3917	0.002147
3	4795	0.6727	0.5015	3.979e-05
4	4838	0.6727	0.503	3.979e-05
5	4774	0.6909	0.5285	1.215e-05
6	4869	0.6909	0.5271	1.215e-05
7	4820	0.6364	0.4508	0.0003385
8	4797	0.6545	0.472	0.0001205
9	4815	0.6182	0.4199	0.0008835
10	4793	0.7273	0.581	8.87e-07
11	4666	0.6	0.3929	0.002147
12	4860	0.6727	0.5065	3.979e-05
13	4713	0.6545	0.4714	0.0001205
14	4667	0.5818	0.3759	0.004866
15	4817	0.6909	0.5387	1.215e-05
16	4844	0.6545	0.4746	0.0001205
17	**4761**	**0.7273**	**0.5831**	**8.87e-07**
18	4850	0.7091	0.5569	3.424e-06
19	4827	0.6727	0.4977	3.979e-05
20	4755	0.6727	0.5043	3.979e-05

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6636	0.04023	0.7273

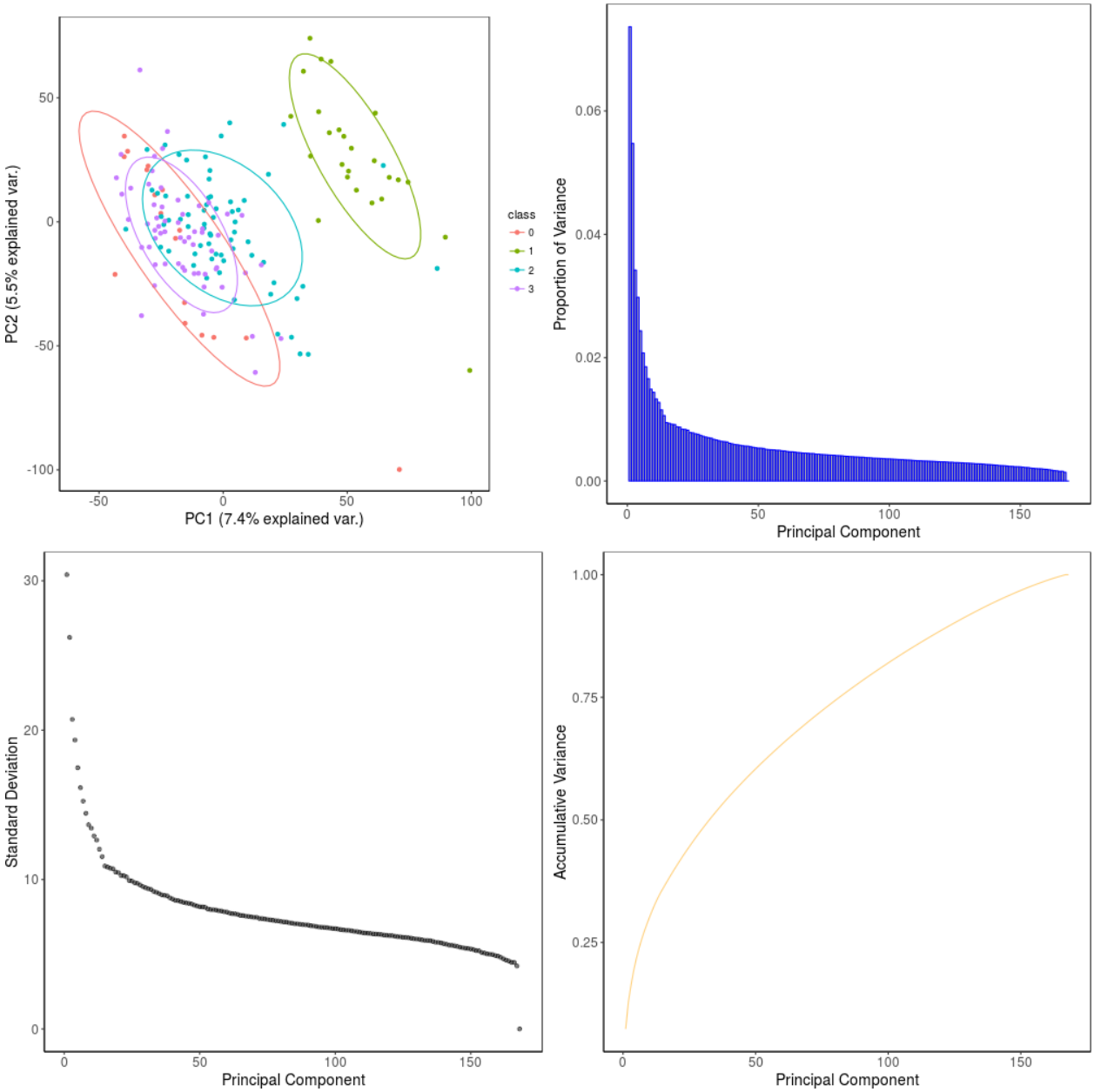
Workflow runtime

7.904 minutes

Plots

Visualization of the classification using PCA

- Groups distribution on the first two Principal Components (PC1 and PC2) from the original data (without apply any FS method).



- Groups distribution on the first two Principal Components (PC1 and PC2) after to apply the FS workflow.