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

Analysis of breast cancer tumor samples using 2-color cDNA microarrays (GSE5325).

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	1843	0.8857	0.7705	6.136e-05
2	1805	0.8571	0.7059	0.0003014
3	1697	0.8571	0.7059	0.0003014
4	1729	0.7429	0.4706	0.02786
5	1916	0.9143	0.8264	9.733e-06
6	1643	0.7714	0.5172	0.01134
7	1695	0.8	0.6016	0.003999
8	1743	0.7429	0.4615	0.02786
9	1964	0.9143	0.8235	9.733e-06
10	1702	0.8286	0.65	0.001202
11	1867	0.9143	0.8264	9.733e-06
12	1885	0.8571	0.6957	0.0003014
13	1923	0.8857	0.7627	6.136e-05
14	1734	0.8286	0.65	0.001202
15	1728	0.8571	0.7059	0.0003014
16	1795	0.8857	0.7627	6.136e-05
17	**1874**	**0.9429**	**0.8852**	**1.128e-06**
18	1743	0.8571	0.7009	0.0003014
19	1679	0.7714	0.541	0.01134
20	1822	0.8571	0.7059	0.0003014

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.8486	0.05796	0.9429

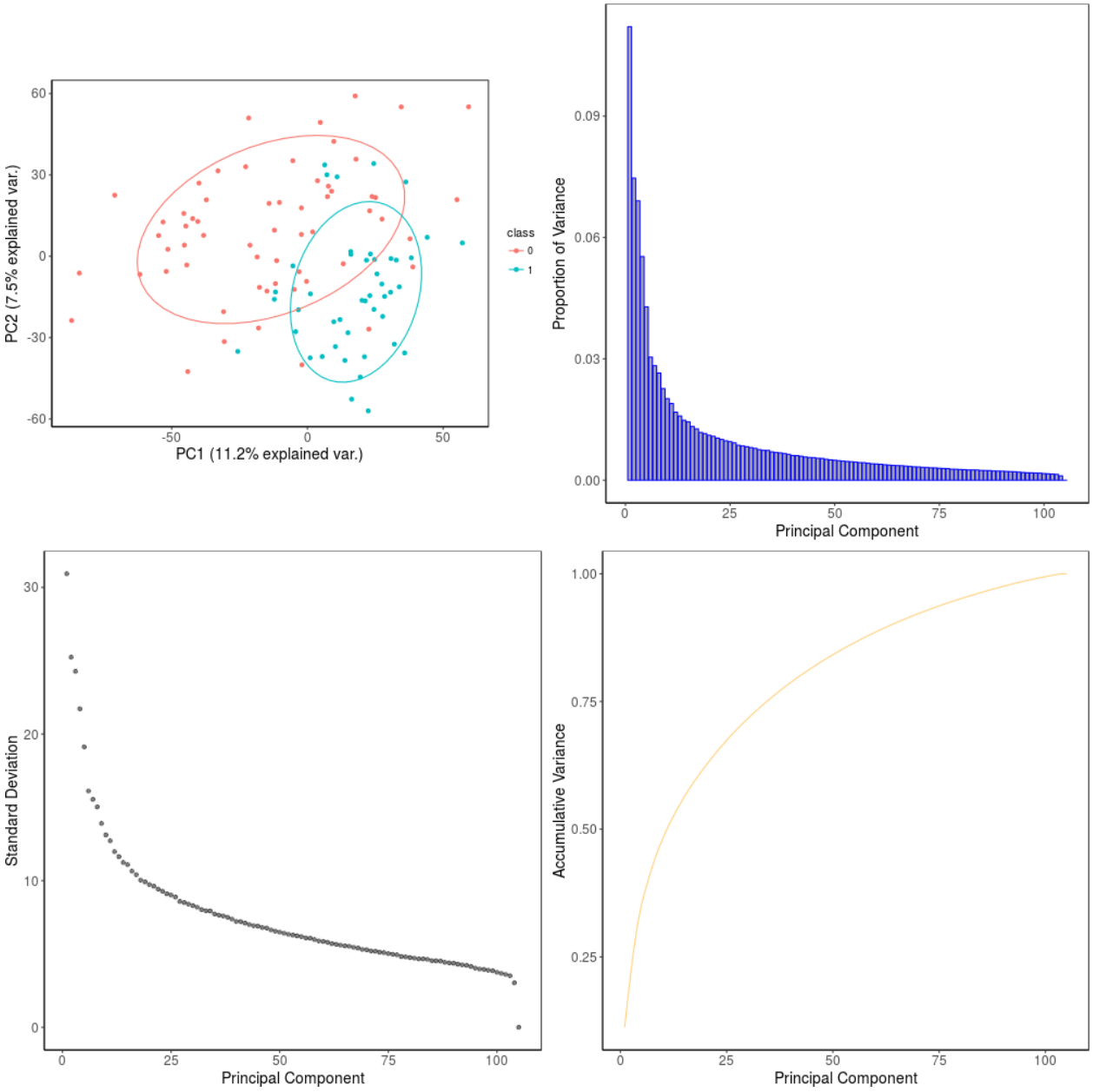
Workflow runtime

2.374 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.