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

Triple-Negative Breast Cancer (TNBC) proteome. Label-free deep proteome analysis of 44 (samples and technical repliques) human breast specimens.

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	1490	1	1	2.216e-07
2	1479	0.8462	0.806	9.42e-05
3	1475	1	1	2.216e-07
4	1484	0.9231	0.9008	6.703e-06
5	1468	0.8462	0.803	9.42e-05
6	1475	0.8462	0.7969	9.42e-05
7	1448	0.7692	0.7111	0.000816
8	1456	0.8462	0.803	9.42e-05
9	1492	1	1	2.216e-07
10	1477	0.8462	0.7969	9.42e-05
11	1439	0.8462	0.7969	9.42e-05
12	1479	0.9231	0.8984	6.703e-06
13	1514	0.9231	0.8984	6.703e-06
14	1457	1	1	2.216e-07
15	1480	0.8462	0.803	9.42e-05
16	**1419**	**1**	**1**	**2.216e-07**
17	1473	1	1	2.216e-07
18	1472	1	1	2.216e-07
19	1464	0.8462	0.806	9.42e-05
20	1515	1	1	2.216e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.9154	0.07853	1

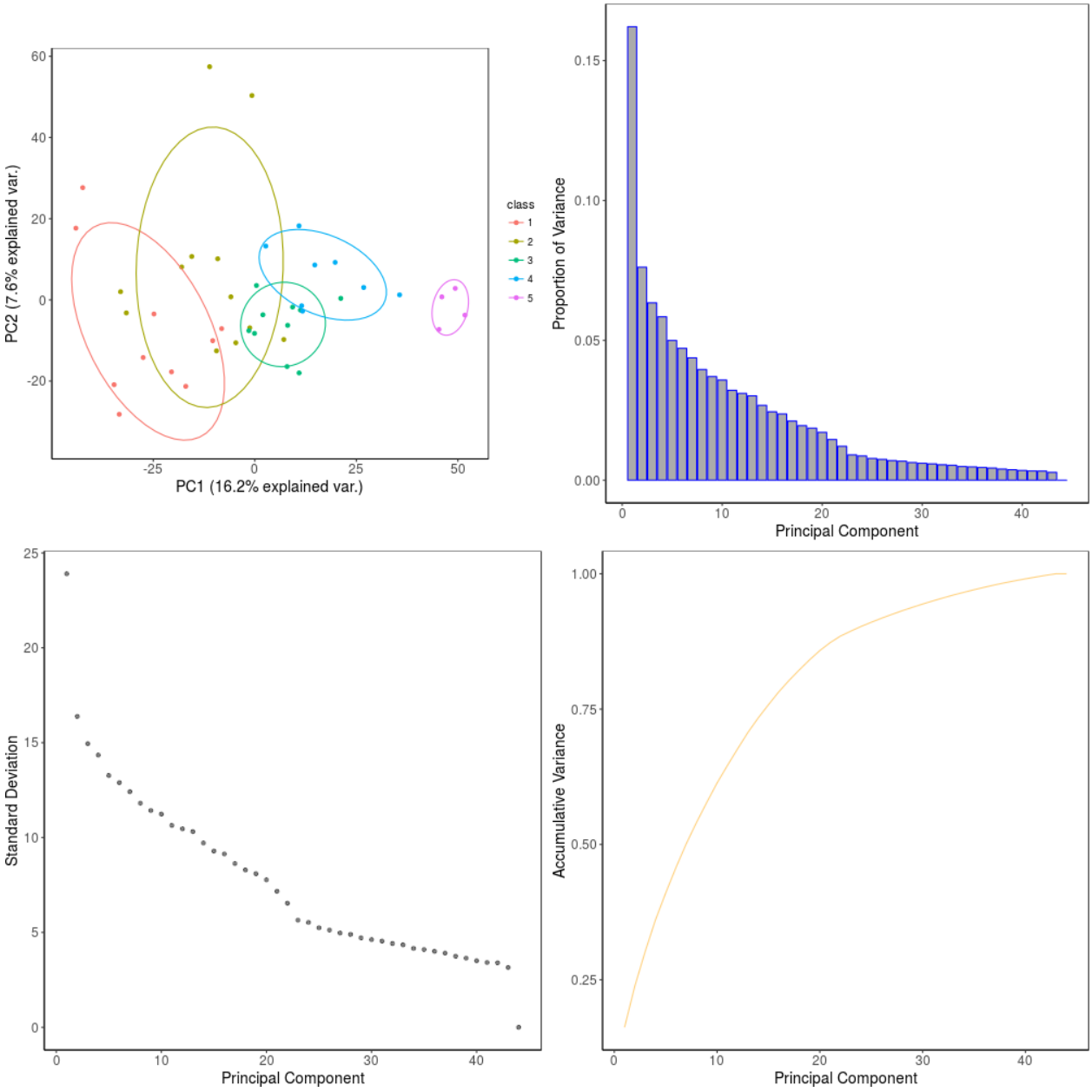
Workflow runtime

0.373 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.