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_GPL93).

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	4999	0.6792	0.5117	2.884e-05
2	4726	0.6415	0.4527	0.0002658
3	4864	0.6792	0.5117	2.884e-05
4	4822	0.6226	0.428	0.0007184
5	4761	0.5849	0.3781	0.004204
6	4775	0.6415	0.453	0.0002658
7	4678	0.6604	0.4857	9.108e-05
8	4929	0.6981	0.5386	8.413e-06
9	4801	0.6415	0.4563	0.0002658
10	4963	0.6604	0.4887	9.108e-05
11	4962	0.6792	0.5071	2.884e-05
12	4716	0.6226	0.4283	0.0007184
13	4680	0.5849	0.3611	0.004204
14	4878	0.6226	0.4252	0.0007184
15	4872	0.6415	0.447	0.0002658
16	4822	0.5849	0.3758	0.004204
17	4808	0.6415	0.4618	0.0002658
18	**4941**	**0.717**	**0.5726**	**2.255e-06**
19	4902	0.6604	0.489	9.108e-05
20	4928	0.6226	0.4252	0.0007184

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6443	0.03636	0.717

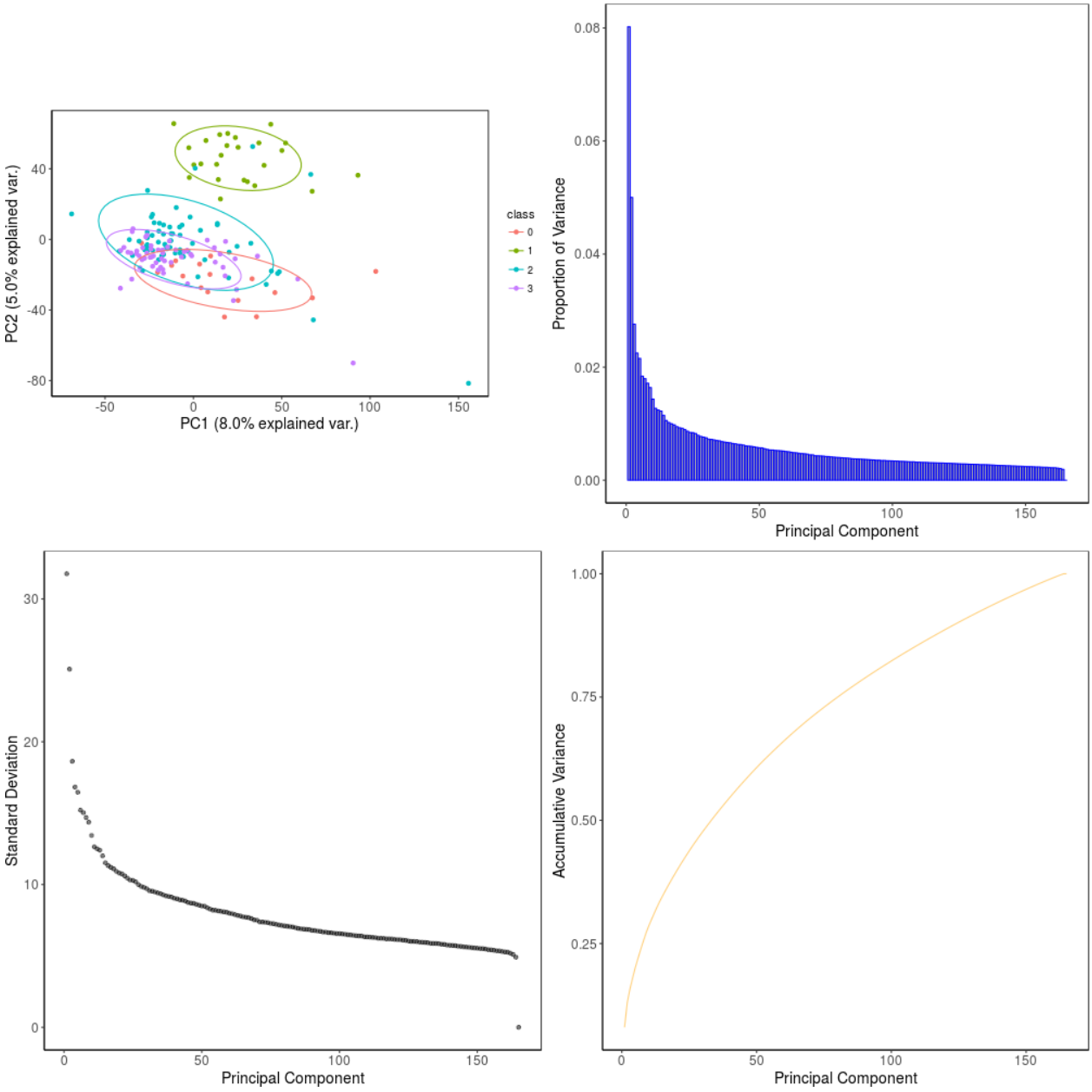
Workflow runtime

7.593 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.