Feature Selection (FS) workflow report

May 18, 2017

Introduction

The report summarizing the Feature Selection pipeline results.

Feature Selection workflow

Univariate canonical correlation (X2) with Multivariate Correlation filter (MC) follow by Recursive Feature Elimination (RFE) wrapped with Random Forest (RF).

Dataset

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

Summary stats from training phase

Table 1: Best model metrics from 10-folds cross-validation resampling.

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.6286	0.2505	0.138	0.2865
2	0.7143	0.4156	0.1166	0.2369
3	0.7143	0.4127	0.1506	0.3073
4	0.7286	0.4408	0.1421	0.2967
5	0.7571	0.4996	0.1355	0.281
6	0.8143	0.624	0.1176	0.2358
7	0.8	0.5907	0.1205	0.2477
8	0.8286	0.6465	0.1127	0.2315
9	0.8	0.5884	0.1205	0.2493
10	0.8286	0.642	0.09035	0.1917
15	0.8	0.5881	0.07377	0.1514
20	0.8	0.5881	0.07377	0.1514
25	0.8	0.5884	0.1205	0.243
30	0.8143	0.6188	0.1355	0.2749
35	0.8143	0.6188	0.1355	0.2749
40	0.8143	0.6188	0.1355	0.2749
45	0.8	0.5884	0.1205	0.243
50	0.8286	0.6492	0.1313	0.2668
60	0.8143	0.6188	0.1355	0.2749
70	0.8286	0.6492	0.1313	0.2668
80	0.8143	0.6188	0.1355	0.2749
90	0.8286	0.6492	0.1313	0.2668
100	0.8143	0.616	0.1355	0.2806
1178	0.8	0.5929	0.1205	0.2399

Summary stats from testing phase

Table 2: Classification metrics from twenty class-balanced and randomized runs.

run	Variables	Accuracy	Kappa	AccuracyPValue
1	1178	0.8571	0.7107	0.0003014
2	8	0.8857	0.7667	6.136e-05
3	9	0.8	0.5882	0.003999
4	7	0.8286	0.6557	0.001202
5	90	0.9429	0.8852	1.128e-06
6	5	0.8857	0.7742	6.136e-05
7	40	0.7429	0.4615	0.02786
8	60	0.8	0.595	0.003999
9	3	0.8286	0.6316	0.001202
10	15	0.9143	0.8264	9.733e-06
11	90	0.8571	0.7107	0.0003014
12	5	0.8571	0.7154	0.0003014
13	60	0.8571	0.7059	0.0003014
14	50	0.7429	0.4615	0.02786
15	2	0.8	0.595	0.003999
16	8	0.8286	0.6613	0.001202
17	**8**	**0.9429**	**0.8833**	**1.128e-06**
18	35	0.9143	0.8264	9.733e-06
19	100	0.8857	0.7667	6.136e-05
20	15	0.8286	0.6613	0.001202

Accuracy__Mean	Accuracy__SD	Accuracy__Max
0.85	0.05705	0.9429

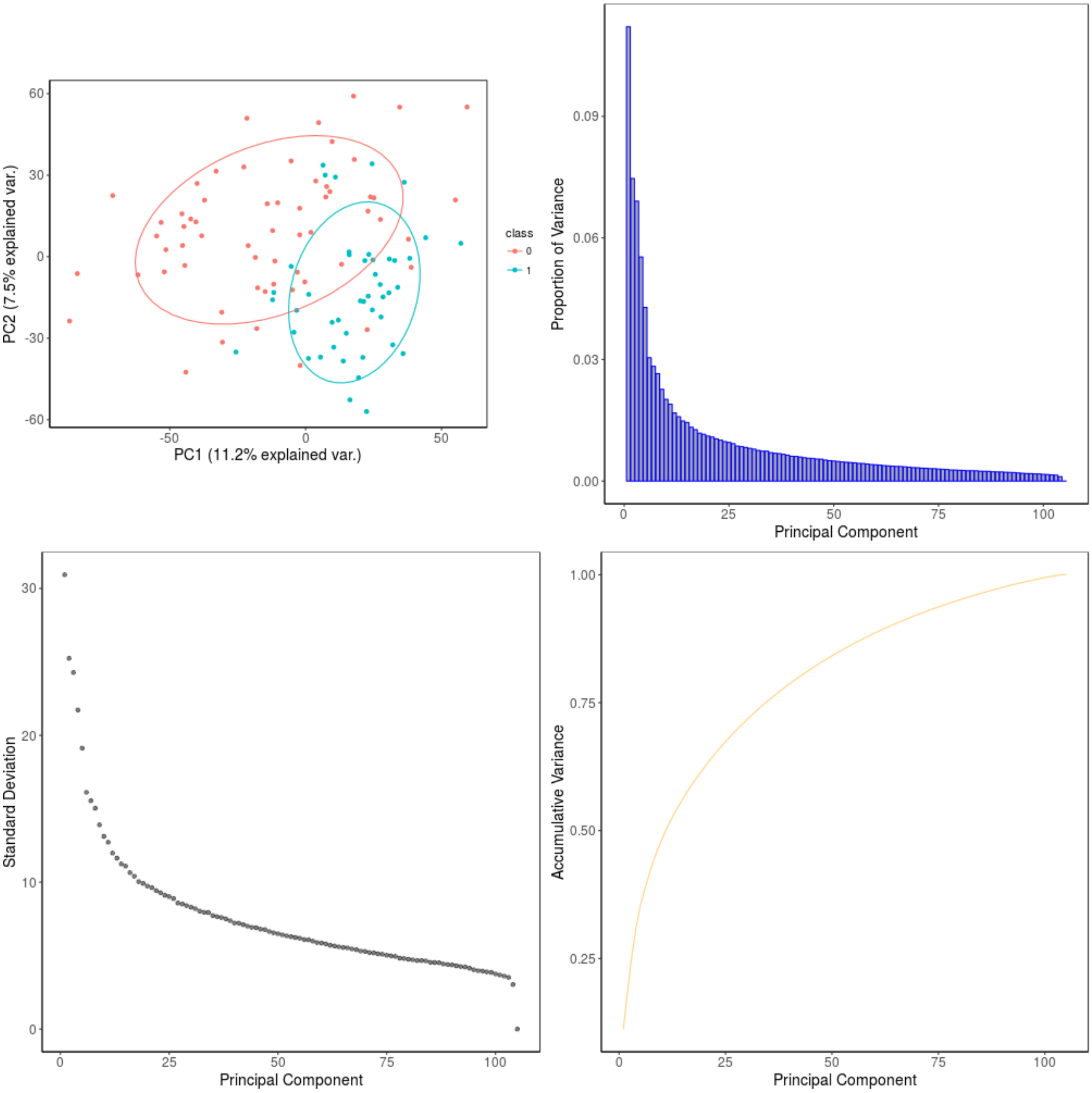
Workflow runtime

6.385 minutes

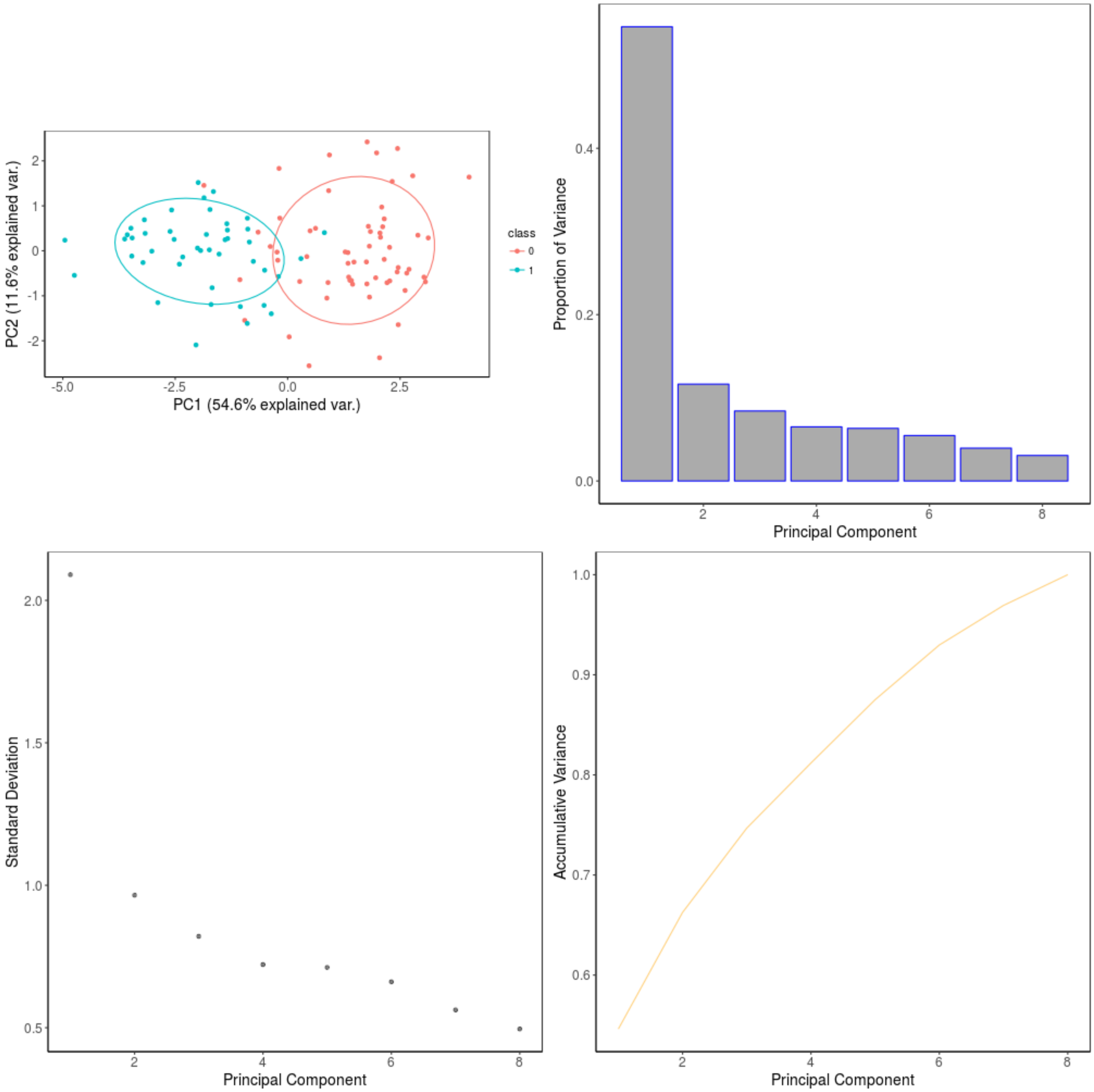
Plots

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.


PC1 (54.6% explained var.)
PC2 (11.6% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component