Feature Selection (FS) workflow report

May 18, 2017

Introduction

The report summarizing the Feature Selection pipeline results.

Feature Selection workflow

Recursive Feature Elimination (RFE) wrapped with Random Forest (RF).

Dataset

Expression data from normal and prostate tumor tissues (GSE6919_GPL8300).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.5159	0.3014	0.1775	0.2566
2	0.5638	0.3739	0.155	0.2105
3	0.5782	0.3863	0.1315	0.1895
4	0.6307	0.4585	0.1995	0.2958
5	0.6488	0.4857	0.1552	0.2263
6	0.6871	0.5417	0.1544	0.2249
7	0.6746	0.5269	0.1331	0.1855
8	0.7198	0.5907	0.1678	0.2444
9	0.7182	0.5922	0.1291	0.1849
10	0.7289	0.6077	0.1496	0.2145
15	0.7156	0.5861	0.1013	0.1463
20	0.7329	0.6037	0.1253	0.1929
25	0.6879	0.5423	0.1681	0.2599
30	0.7046	0.5642	0.1139	0.1808
35	0.7307	0.6057	0.1334	0.2024
40	0.7146	0.58	0.1353	0.2114
45	0.6788	0.5248	0.1548	0.2445
50	0.6946	0.5486	0.1437	0.2306
60	0.7106	0.5718	0.1465	0.2347
70	0.7123	0.5763	0.1253	0.1977
80	0.7464	0.6278	0.1218	0.1916
90	0.7023	0.5631	0.1333	0.2136
100	0.7098	0.572	0.1473	0.2354
12558	0.6056	0.4048	0.1379	0.2169

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	60	0.6786	0.5348	4.002e-06
2	80	0.6429	0.4872	4.381e-05
3	50	0.6429	0.4805	4.381e-05
4	100	0.7143	0.5865	2.677e-07
5	70	0.7321	0.6127	6.116e-08
6	25	0.75	0.6407	1.281e-08
7	90	0.6964	0.5637	1.078e-06
8	60	0.7143	0.5922	2.677e-07
9	50	0.6429	0.4897	4.381e-05
10	80	0.75	0.6312	1.281e-08
11	70	0.6429	0.4732	4.381e-05
12	100	0.6607	0.506	1.375e-05
13	15	0.6429	0.4796	4.381e-05
14	90	0.75	0.6428	1.281e-08
15	80	0.75	0.6335	1.281e-08
16	90	0.7679	0.6573	2.45e-09
17	90	0.7857	0.688	4.256e-10
18	**80**	**0.7857**	**0.6902**	**4.256e-10**
19	30	0.6607	0.509	1.375e-05
20	80	0.6964	0.5645	1.078e-06

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.7054	0.05133	0.7857

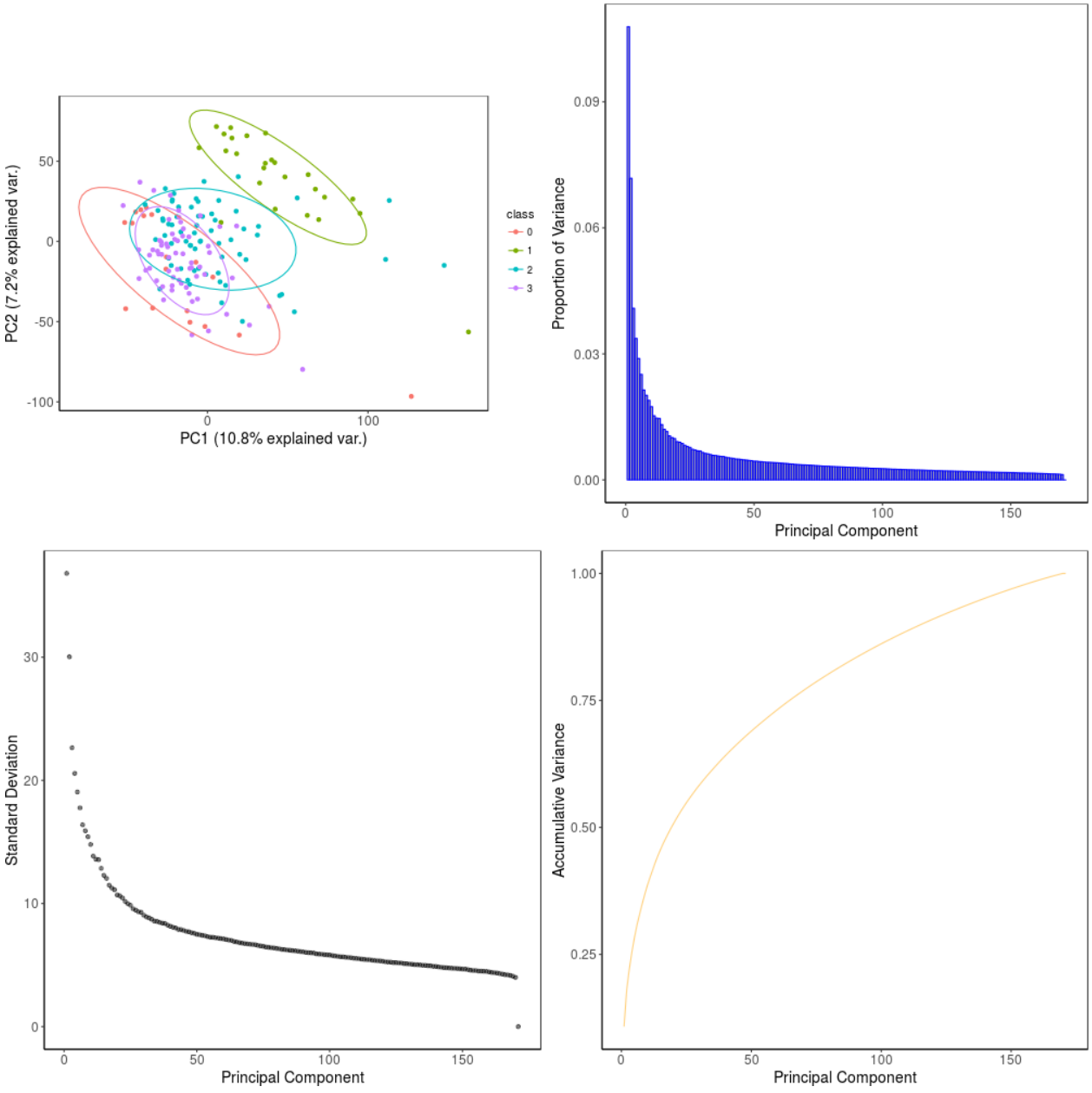
Workflow runtime

74.553 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.