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

Expression data from normal and prostate tumor tissues (GSE6919_GPL93).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4553	0.208	0.1297	0.1822
2	0.5061	0.2678	0.09259	0.1535
3	0.5654	0.3547	0.1292	0.2117
4	0.5576	0.3413	0.1028	0.1608
5	0.6018	0.4033	0.1006	0.1717
6	0.5873	0.3829	0.09309	0.1501
7	0.5762	0.3653	0.1034	0.1724
8	0.5684	0.3577	0.09698	0.1601
9	0.5948	0.3999	0.09452	0.158
10	0.5669	0.3597	0.08491	0.1386
15	0.5396	0.3116	0.1263	0.1936
20	0.5623	0.3446	0.1269	0.2078
25	0.638	0.4602	0.1189	0.1801
30	0.6052	0.415	0.11	0.1648
35	0.6237	0.4477	0.1118	0.1584
40	0.6502	0.4791	0.1162	0.1706
45	0.6616	0.503	0.1195	0.1739
50	0.6954	0.5471	0.105	0.1645
60	0.6841	0.5314	0.09131	0.1379
70	0.6904	0.5355	0.1151	0.1932
80	0.6841	0.5281	0.08243	0.128
90	0.6778	0.5212	0.06956	0.09924
100	0.6674	0.5002	0.06359	0.1111
459	0.6788	0.5205	0.08209	0.1198

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	35	0.7358	0.6099	5.533e-07
2	459	0.6792	0.5148	2.884e-05
3	45	0.6226	0.4329	0.0007184
4	80	0.6981	0.5504	8.413e-06
5	70	0.7358	0.6066	5.533e-07
6	40	0.6604	0.4871	9.108e-05
7	459	0.6604	0.486	9.108e-05
8	70	0.6981	0.5361	8.413e-06
9	459	0.6604	0.4826	9.108e-05
10	60	0.717	0.5789	2.255e-06
11	100	0.6792	0.5225	2.884e-05
12	90	0.717	0.5771	2.255e-06
13	459	0.7358	0.6024	5.533e-07
14	459	0.7358	0.5959	5.533e-07
15	90	0.717	0.5803	2.255e-06
16	10	0.717	0.578	2.255e-06
17	70	0.6604	0.4832	9.108e-05
18	90	0.6415	0.4606	0.0002658
19	**50**	**0.7547**	**0.6284**	**1.238e-07**
20	100	0.6226	0.4307	0.0007184

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6925	0.04019	0.7547

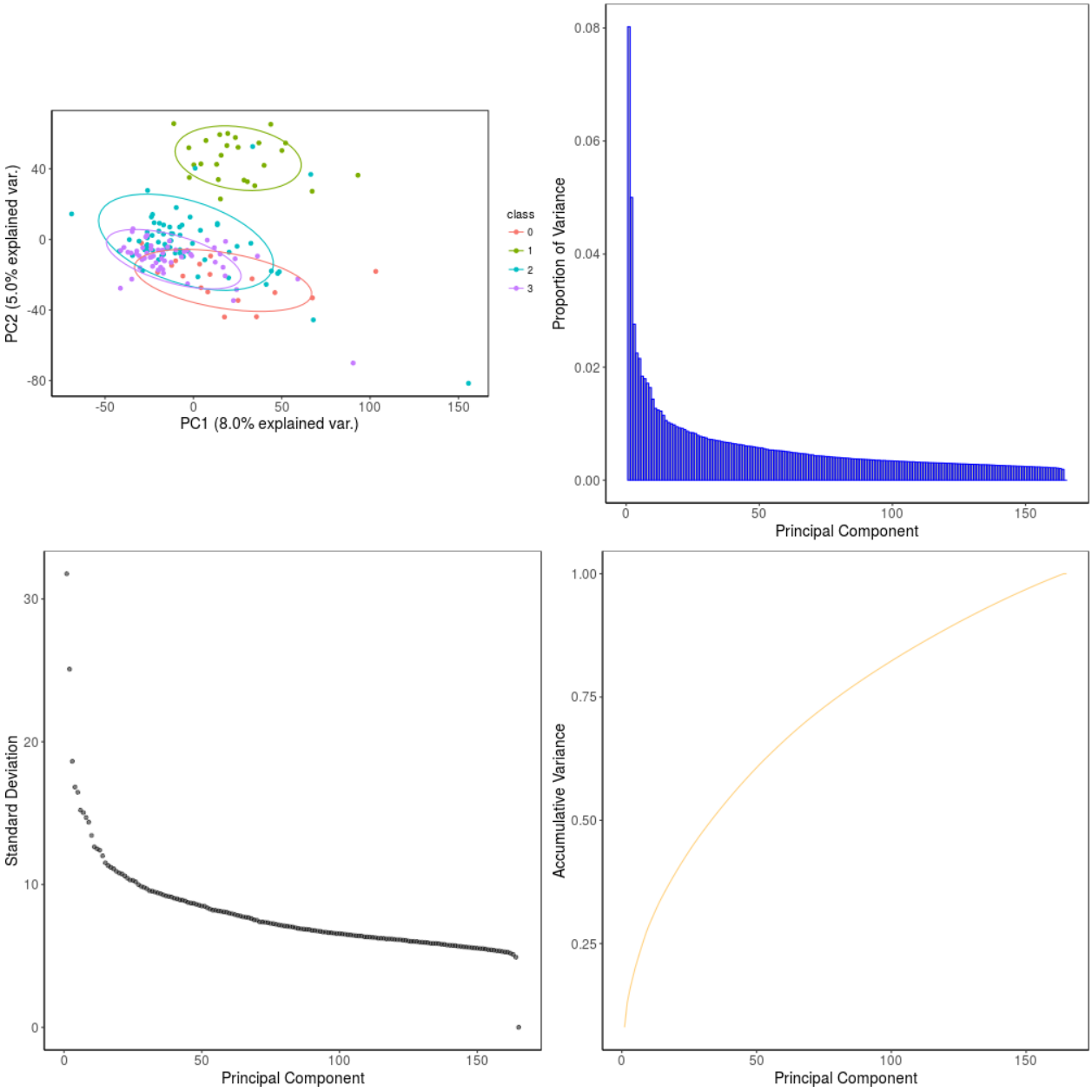
Workflow runtime

8.664 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.