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_GPL8300).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.5584	0.3525	0.1223	0.165
2	0.4863	0.2349	0.1062	0.17
3	0.5305	0.3069	0.1674	0.2532
4	0.5235	0.2879	0.1565	0.2329
5	0.5591	0.3379	0.1273	0.1922
6	0.5703	0.3482	0.1361	0.2158
7	0.5863	0.3779	0.1014	0.1541
8	0.5876	0.3824	0.1182	0.1671
9	0.5862	0.3831	0.1064	0.1513
10	0.5612	0.341	0.1144	0.1709
15	0.6015	0.3971	0.136	0.2083
20	0.6568	0.4783	0.1345	0.2087
25	0.6363	0.4502	0.1168	0.1775
30	0.6545	0.4817	0.1096	0.1685
35	0.6445	0.4657	0.1653	0.2565
40	0.653	0.4783	0.1276	0.198
45	0.669	0.502	0.13	0.201
50	0.6886	0.5314	0.1306	0.1977
60	0.6257	0.442	0.1247	0.1795
70	0.6691	0.5065	0.08924	0.1301
80	0.6608	0.492	0.06148	0.09258
90	0.6448	0.4676	0.07477	0.1111
100	0.6706	0.5072	0.08379	0.1266
349	0.6796	0.5201	0.09817	0.1474

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	45	0.6786	0.5186	4.002e-06
2	50	0.7143	0.5858	2.677e-07
3	90	0.7143	0.5748	2.677e-07
4	**50**	**0.8393**	**0.7658**	**1.182e-12**
5	40	0.7679	0.652	2.45e-09
6	50	0.6964	0.5545	1.078e-06
7	90	0.7321	0.6093	6.116e-08
8	70	0.6786	0.5156	4.002e-06
9	60	0.6964	0.5572	1.078e-06
10	60	0.625	0.4308	0.0001297
11	80	0.7321	0.6069	6.116e-08
12	35	0.6429	0.4786	4.381e-05
13	40	0.7143	0.5833	2.677e-07
14	20	0.7143	0.5833	2.677e-07
15	20	0.7143	0.5782	2.677e-07
16	100	0.6429	0.4791	4.381e-05
17	90	0.6964	0.558	1.078e-06
18	80	0.7857	0.6791	4.256e-10
19	70	0.6607	0.509	1.375e-05
20	100	0.7143	0.5807	2.677e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.708	0.05027	0.8393

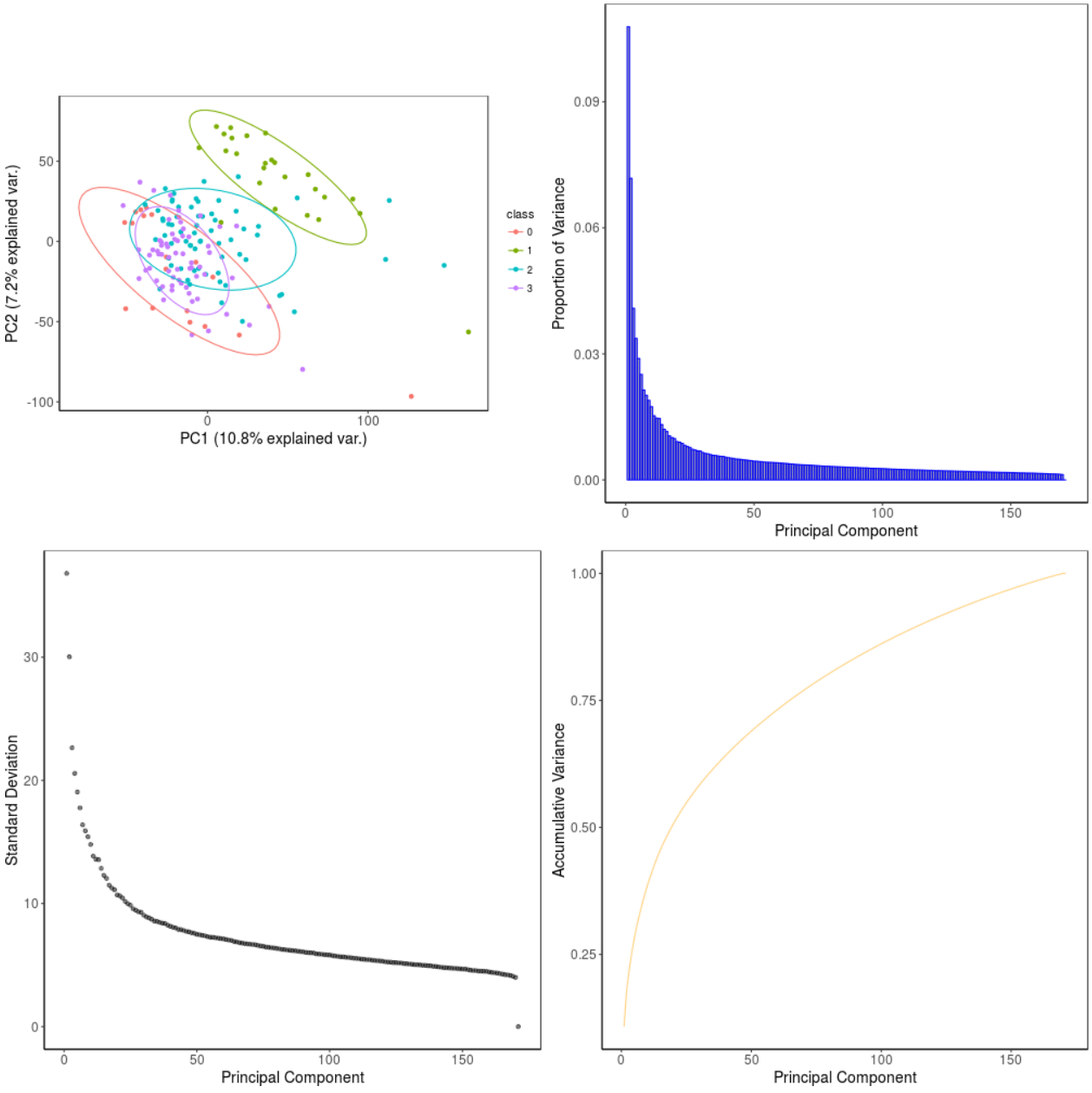
Workflow runtime

8.269 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.