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_GPL92).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4633	0.212	0.1697	0.24
2	0.4823	0.2361	0.2095	0.3043
3	0.4912	0.2492	0.1436	0.216
4	0.5162	0.2807	0.1066	0.1668
5	0.5567	0.3435	0.1254	0.1827
6	0.5359	0.2982	0.1691	0.2629
7	0.5452	0.3185	0.1057	0.167
8	0.58	0.3685	0.1295	0.2048
9	0.59	0.3832	0.1209	0.1871
10	0.5983	0.3953	0.11	0.1775
15	0.6003	0.3968	0.09839	0.1529
20	0.6561	0.4836	0.1232	0.1835
25	0.6627	0.4909	0.1412	0.2164
30	0.6411	0.4612	0.1278	0.1897
35	0.6203	0.4272	0.09142	0.138
40	0.5918	0.3795	0.07522	0.1311
45	0.6118	0.413	0.1001	0.158
50	0.5636	0.3371	0.1061	0.1738
60	0.6085	0.4072	0.08111	0.1403
70	0.6426	0.4591	0.1306	0.2073
80	0.6376	0.4522	0.1141	0.1766
90	0.655	0.4814	0.113	0.1768
100	0.5992	0.3923	0.09667	0.1658
12553	0.6268	0.435	0.1154	0.1677

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	30	0.7273	0.5952	8.87e-07
2	50	0.6727	0.5094	3.979e-05
3	60	0.6545	0.4754	0.0001205
4	100	0.6	0.395	0.002147
5	**25**	**0.7273**	**0.5838**	**8.87e-07**
6	35	0.6364	0.4516	0.0003385
7	90	0.6364	0.4514	0.0003385
8	70	0.7091	0.5611	3.424e-06
9	100	0.7273	0.5936	8.87e-07
10	60	0.6545	0.4728	0.0001205
11	30	0.6545	0.4738	0.0001205
12	80	0.5455	0.3156	0.02042
13	20	0.6	0.4048	0.002147
14	70	0.6545	0.4759	0.0001205
15	90	0.6545	0.4749	0.0001205
16	50	0.6909	0.5441	1.215e-05
17	80	0.6909	0.5273	1.215e-05
18	25	0.7091	0.5628	3.424e-06
19	20	0.6909	0.5327	1.215e-05
20	30	0.6909	0.5463	1.215e-05

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6664	0.04766	0.7273

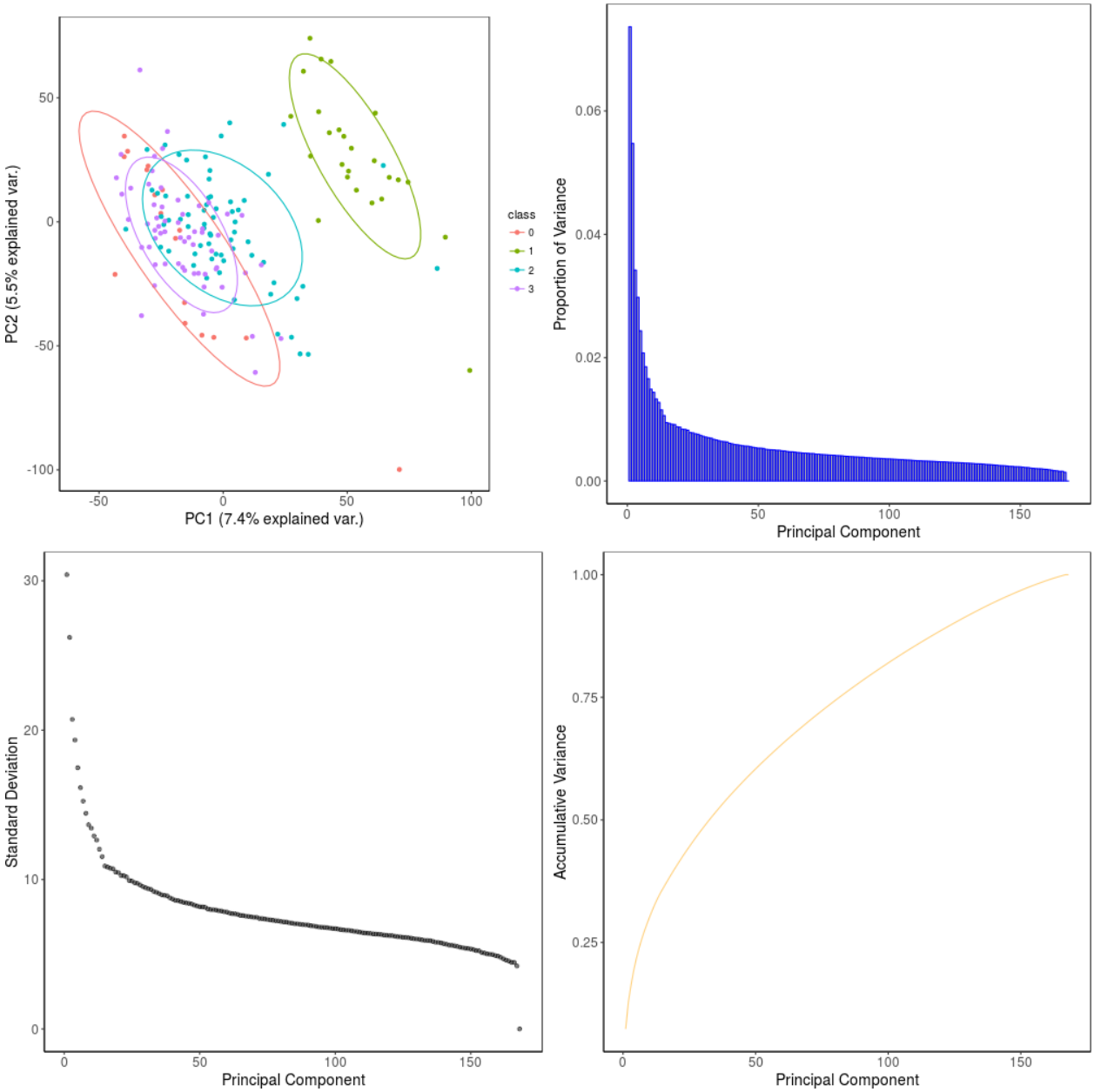
Workflow runtime

85.871 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.