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

Transcriptomics analysis of left ventricles of mouse subjected to Isoproterenol challenge (GSE48760).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.7284	0.4564	0.1339	0.2676
2	0.7856	0.5711	0.1428	0.2854
3	0.8219	0.6439	0.123	0.2461
4	0.8071	0.6139	0.1306	0.2609
5	0.829	0.6582	0.14	0.2801
6	0.8428	0.6853	0.1156	0.2311
7	0.829	0.6576	0.1174	0.2345
8	0.8218	0.6433	0.1224	0.2445
9	0.8218	0.6433	0.1177	0.235
10	0.8361	0.6718	0.1117	0.2231
15	0.8218	0.6433	0.1128	0.2252
20	0.8284	0.6558	0.09649	0.1928
25	0.8367	0.6732	0.1219	0.2437
30	0.829	0.6576	0.1019	0.2035
35	0.8438	0.6875	0.1108	0.2216
40	0.8295	0.6588	0.101	0.2016
45	0.8438	0.688	0.1108	0.2218
50	0.8438	0.6875	0.1108	0.2216
60	0.8295	0.6588	0.101	0.2016
70	0.8361	0.6723	0.101	0.2019
80	0.8428	0.6853	0.1204	0.2407
90	0.8285	0.6568	0.1223	0.2444
100	0.8214	0.6425	0.1357	0.2712
25697	0.8295	0.6588	0.101	0.2016

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	5	0.8986	0.7971	4.536e-12
2	9	0.8841	0.7681	3.485e-11
3	80	0.8841	0.7683	3.485e-11
4	30	0.8551	0.7101	1.398e-09
5	90	0.8696	0.7392	2.345e-10
6	4	0.8116	0.6239	1.572e-07
7	35	0.8841	0.7683	3.485e-11
8	8	0.9275	0.8551	4.82e-14
9	30	0.8986	0.797	4.536e-12
10	9	0.913	0.8261	5.088e-13
11	20	0.913	0.8263	5.088e-13
12	7	0.8551	0.7106	1.398e-09
13	10	0.8551	0.7098	1.398e-09
14	50	0.8696	0.7394	2.345e-10
15	40	0.8696	0.739	2.345e-10
16	**35**	**0.942**	**0.884**	**3.75e-15**
17	90	0.9275	0.855	4.82e-14
18	6	0.8551	0.7106	1.398e-09
19	35	0.9275	0.855	4.82e-14
20	60	0.8841	0.7685	3.485e-11

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.8862	0.03267	0.942

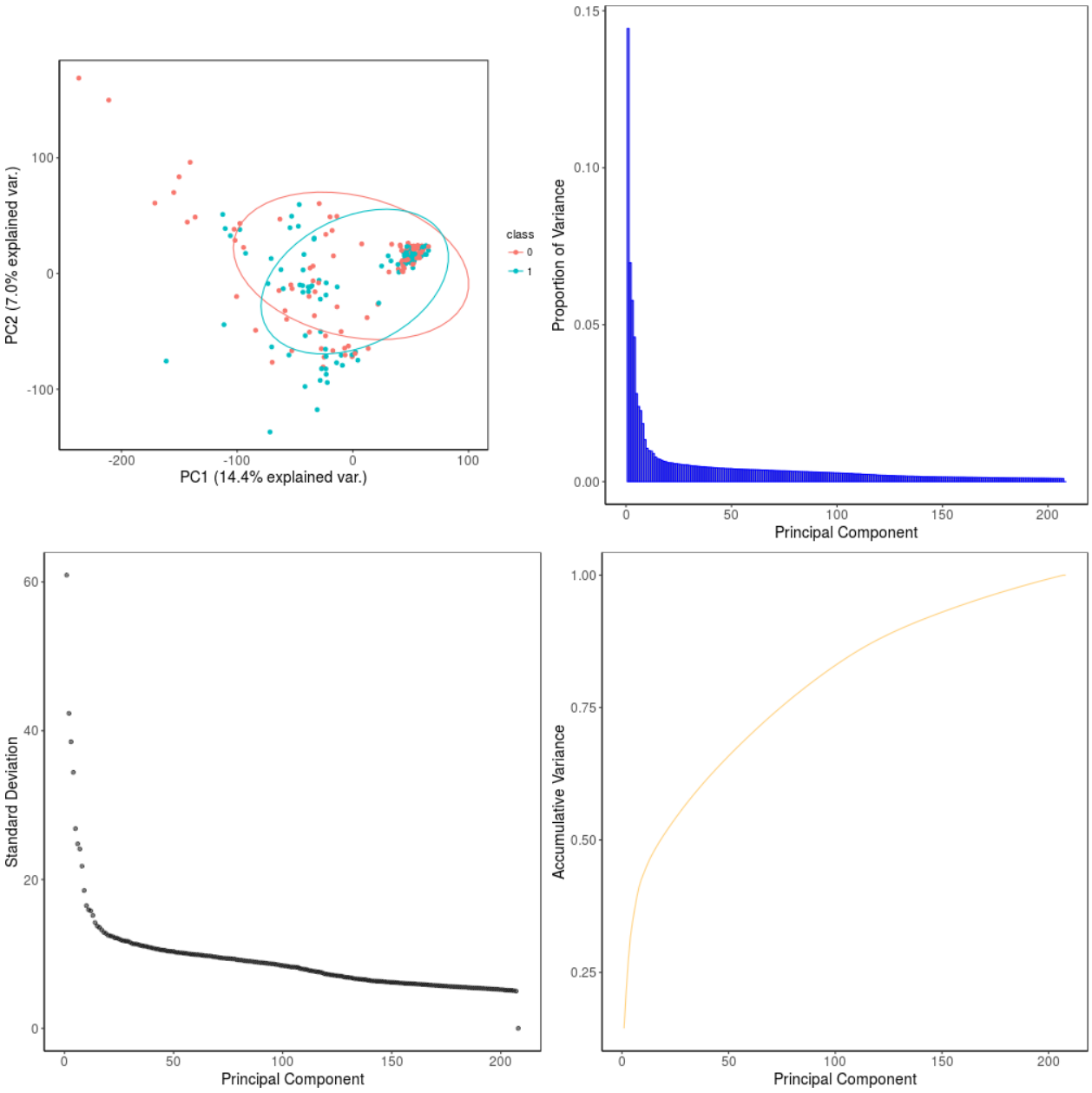
Workflow runtime

136.857 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.