

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

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	0.7338	0.4673	0.1471	0.2946
2	0.8473	0.6953	0.1047	0.2092
3	0.8324	0.6664	0.1033	0.2056
4	0.8396	0.6807	0.1076	0.2141
5	0.833	0.6671	0.0964	0.1918
6	0.8401	0.6814	0.1213	0.2418
7	0.8692	0.7388	0.1003	0.2006
8	0.8621	0.7245	0.1142	0.2283
9	0.8401	0.6814	0.1346	0.2685
10	0.8544	0.7099	0.1179	0.2348
15	0.8473	0.6956	0.11	0.219
20	0.8473	0.6956	0.1333	0.2658
25	0.8544	0.7099	0.1179	0.2348
30	0.8544	0.7099	0.1179	0.2348
35	0.8544	0.7099	0.1179	0.2348
40	0.8544	0.7099	0.1179	0.2348
45	0.8544	0.7099	0.1179	0.2348
50	0.8544	0.7099	0.1179	0.2348
60	0.8615	0.7242	0.1051	0.2091
70	0.8615	0.7242	0.1051	0.2091
80	0.8687	0.7385	0.1124	0.2235
90	0.8687	0.7385	0.1124	0.2235
100	0.8615	0.7242	0.1249	0.2487
482	0.8687	0.7385	0.1124	0.2235

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	7	0.9275	0.8551	4.82e-14
2	40	0.8986	0.7968	4.536e-12
3	6	0.8696	0.739	2.345e-10
4	482	0.8986	0.7971	4.536e-12
5	7	0.8841	0.7685	3.485e-11
6	20	0.8841	0.7683	3.485e-11
7	25	0.8406	0.682	7.46e-09
8	6	0.913	0.8259	5.088e-13
9	80	0.8116	0.6239	1.572e-07
10	60	0.9275	0.8549	4.82e-14
11	9	0.8261	0.6524	3.592e-08
12	80	0.8551	0.7103	1.398e-09
13	15	0.8986	0.7973	4.536e-12
14	40	0.8696	0.7396	2.345e-10
15	9	0.8551	0.7101	1.398e-09
16	40	0.8406	0.6815	7.46e-09
17	80	0.913	0.8261	5.088e-13
18	482	0.913	0.8262	5.088e-13
19	50	0.8696	0.7398	2.345e-10
20	482	0.913	0.8259	5.088e-13

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.8804	0.03419	0.9275

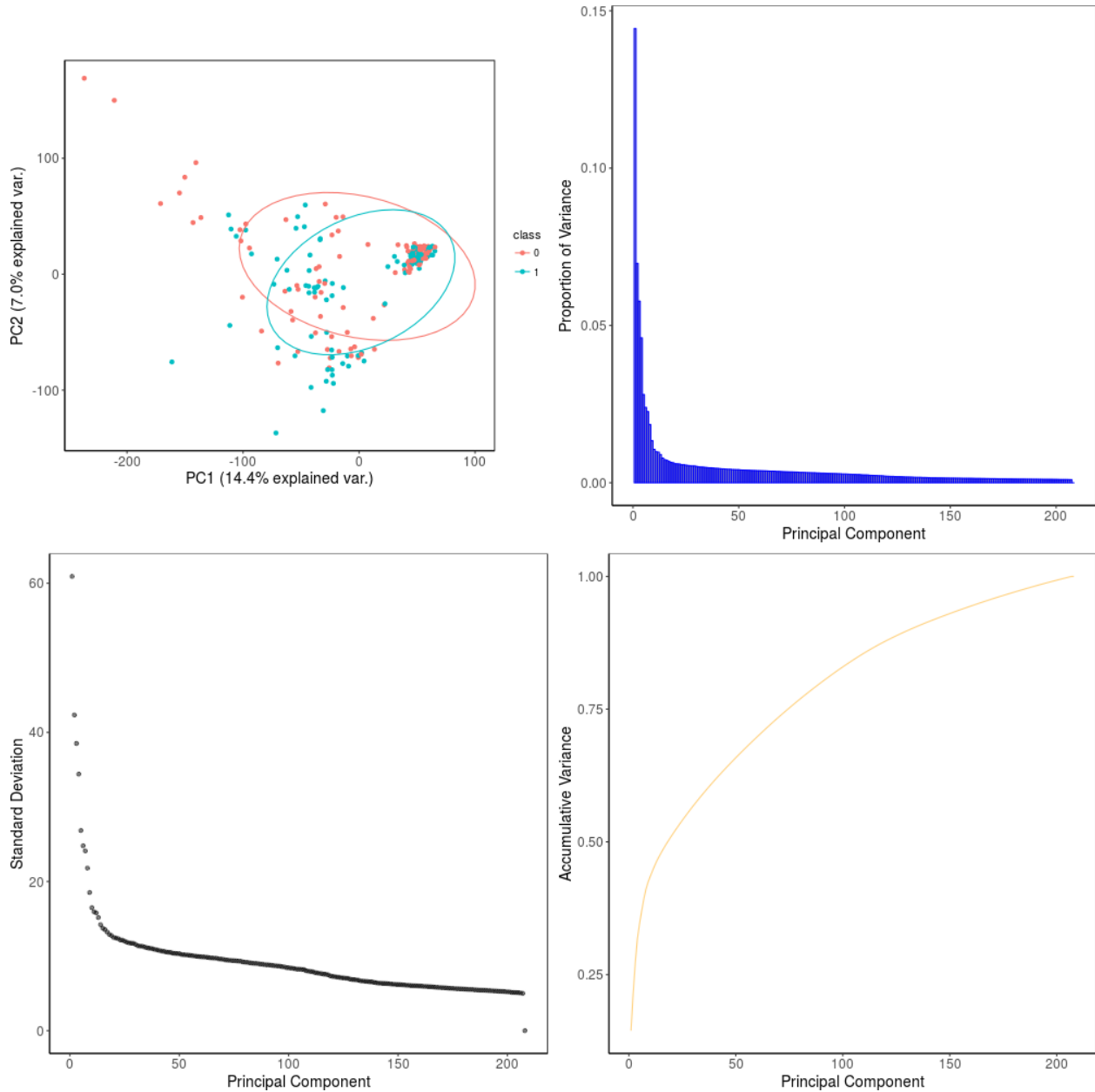
Workflow runtime

8.885 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.

