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

Analysis of breast cancer tumor samples using 2-color cDNA microarrays (GSE5325).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.6143	0.2239	0.2524	0.4908
2	0.7286	0.4331	0.1421	0.2952
3	0.7286	0.4409	0.2177	0.4329
4	0.7714	0.525	0.2043	0.4221
5	0.8286	0.6468	0.1622	0.3363
6	0.8286	0.6493	0.1998	0.407
7	0.8143	0.6188	0.1911	0.3888
8	0.8143	0.6188	0.1911	0.3888
9	0.8143	0.6188	0.1911	0.3888
10	0.8143	0.6185	0.1656	0.3392
15	0.8571	0.7075	0.1347	0.2795
20	0.8571	0.7048	0.1506	0.3174
25	0.8571	0.7048	0.1506	0.3174
30	0.8714	0.738	0.1421	0.2942
35	0.8571	0.7048	0.1506	0.3174
40	0.8571	0.7048	0.1506	0.3174
45	0.8714	0.738	0.1421	0.2942
50	0.8714	0.738	0.1421	0.2942
60	0.8429	0.6794	0.1838	0.3768
70	0.8571	0.7126	0.1782	0.3601
80	0.8429	0.6794	0.1838	0.3768
90	0.8571	0.7048	0.1506	0.3174
100	0.8429	0.6794	0.1838	0.3768
8534	0.8429	0.6684	0.171	0.3762

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	4	0.8286	0.6441	0.001202
2	6	0.8	0.5812	0.003999
3	20	0.8571	0.7059	0.0003014
4	45	0.8286	0.6379	0.001202
5	6	0.8286	0.6613	0.001202
6	25	0.8857	0.7705	6.136e-05
7	15	0.8857	0.7667	6.136e-05
8	8	0.8571	0.7154	0.0003014
9	80	0.8571	0.7009	0.0003014
10	3	0.8286	0.6557	0.001202
11	70	0.7714	0.5484	0.01134
12	6	0.8286	0.6316	0.001202
13	6	0.7714	0.541	0.01134
14	100	0.7714	0.541	0.01134
15	60	0.8571	0.7059	0.0003014
16	10	0.7714	0.5556	0.01134
17	8534	0.8	0.5812	0.003999
18	6	0.8571	0.7154	0.0003014
19	100	0.8	0.6016	0.003999
20	**30**	**0.9143**	**0.8264**	**9.733e-06**

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.83	0.04195	0.9143

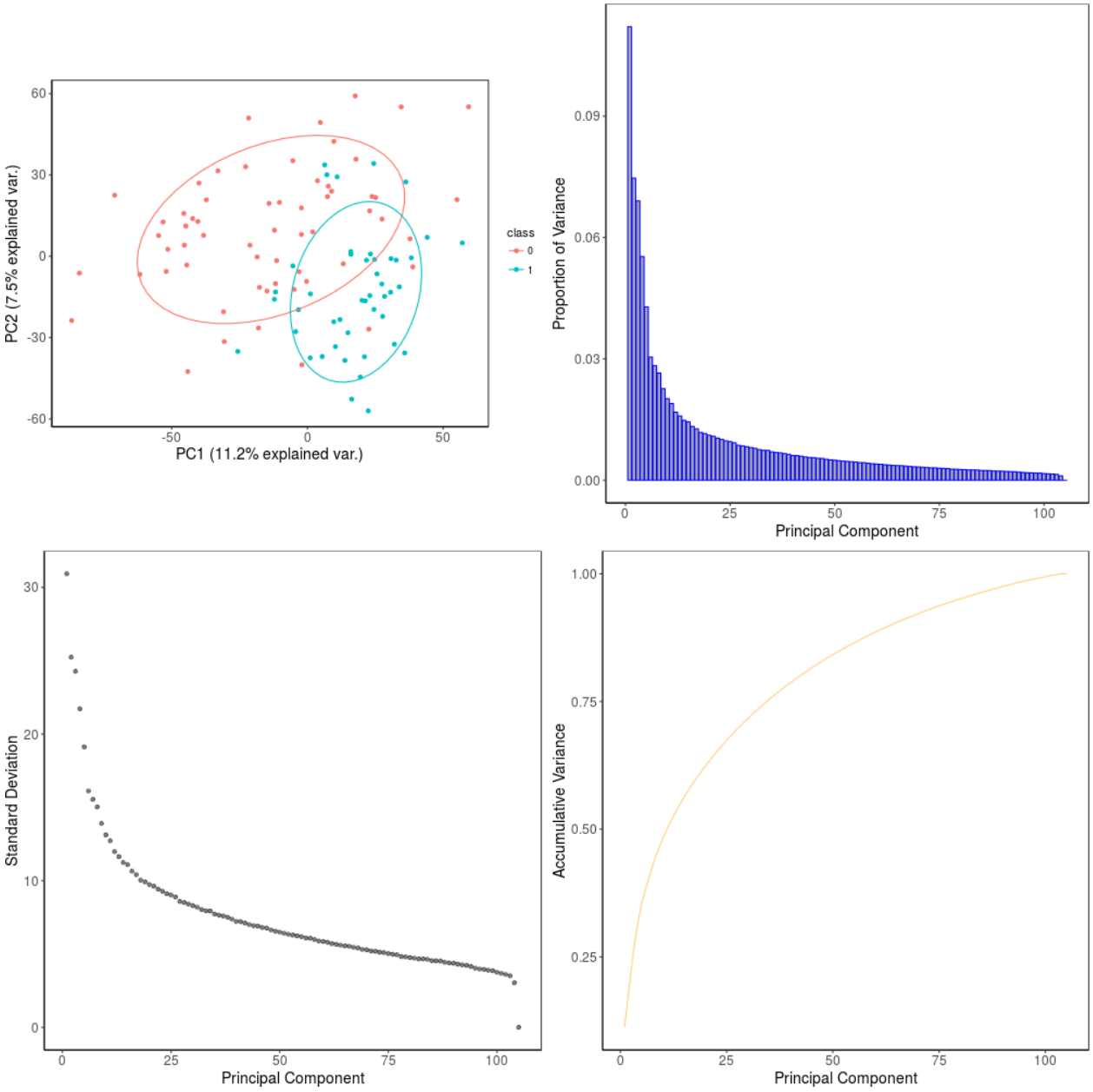
Workflow runtime

25.028 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.