Feature Selection (FS) workflow report

May 18, 2017

Introduction

The report summarizing the Feature Selection pipeline results.

Feature Selection workflow

Principal Component Analysis (PCA) 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	0.4199	0.1468	0.1313	0.2176
2	0.4729	0.2312	0.1343	0.1931
3	0.5934	0.3957	0.096	0.1315
4	0.5882	0.3932	0.1693	0.2553
5	0.6388	0.4643	0.1642	0.2371
6	0.5834	0.3792	0.1726	0.2562
7	0.622	0.4367	0.2287	0.3417
8	0.5788	0.3684	0.1787	0.2627
9	0.6001	0.3951	0.1529	0.2267
10	0.5827	0.3727	0.13	0.1933
15	0.601	0.3968	0.1757	0.2609
20	0.6102	0.4143	0.1709	0.2532
25	0.5646	0.3414	0.1555	0.2334
30	0.5312	0.293	0.1381	0.1997
35	0.5472	0.3162	0.1191	0.1722
40	0.5481	0.3152	0.1652	0.2382
45	0.5486	0.3167	0.1185	0.1682
50	0.5737	0.354	0.1278	0.1822
60	0.5046	0.2474	0.153	0.226
70	0.5148	0.2591	0.145	0.2099
80	0.5408	0.2996	0.1662	0.2496
90	0.5148	0.2588	0.114	0.1727
100	0.5664	0.3366	0.1293	0.1875
171	0.4775	0.1913	0.1292	0.1982

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	5	0.6429	0.48	4.381e-05
2	8	0.7321	0.5989	6.116e-08
3	40	0.6071	0.4111	0.0003574
4	8	0.6071	0.4164	0.0003574
5	8	0.6429	0.4764	4.381e-05
6	8	0.6786	0.5288	4.002e-06
7	4	0.5536	0.3708	0.004947
8	6	0.6607	0.495	1.375e-05
9	20	0.7143	0.5725	2.677e-07
10	25	0.5714	0.3535	0.002204
11	7	0.5357	0.3151	0.0104
12	10	0.625	0.4424	0.0001297
13	70	0.5714	0.3447	0.002204
14	6	0.6071	0.408	0.0003574
15	6	0.6429	0.4832	4.381e-05
16	7	0.6786	0.5186	4.002e-06
17	70	0.5536	0.326	0.004947
18	**5**	**0.7679**	**0.6644**	**2.45e-09**
19	15	0.625	0.4343	0.0001297
20	10	0.6429	0.4764	4.381e-05

Accuracy__Mean	Accuracy__SD	Accuracy__Max
0.633	0.06144	0.7679

Workflow runtime

8.191 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).

```
## PCA plot not available for this FS workflow setting
```

- Groups distribution on the first two Principal Components (PC1 and PC2) after to apply the FS workflow.

```
## PCA plot not available for this FS workflow setting
```