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 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	0.6135	0.4221	0.09185	0.1321
2	0.6161	0.4294	0.1087	0.1592
3	0.6032	0.4093	0.1475	0.2238
4	0.6329	0.4596	0.1332	0.1955
5	0.6243	0.4408	0.1126	0.1689
6	0.7121	0.573	0.1087	0.1611
7	0.6925	0.5449	0.1342	0.199
8	0.6908	0.5409	0.1648	0.2459
9	0.7173	0.5785	0.1378	0.2052
10	0.6908	0.5414	0.1807	0.2688
15	0.7014	0.556	0.1521	0.2242
20	0.6937	0.5423	0.1574	0.2326
25	0.6869	0.5328	0.1444	0.2104
30	0.7114	0.5677	0.1375	0.2036
35	0.7212	0.5838	0.153	0.2243
40	0.6912	0.5372	0.1639	0.2422
45	0.6662	0.4994	0.1919	0.2852
50	0.6659	0.4925	0.1849	0.2777
60	0.6572	0.4849	0.1755	0.2624
70	0.639	0.4552	0.153	0.2278
80	0.6238	0.428	0.1674	0.2516
90	0.5993	0.389	0.1468	0.2193
100	0.6193	0.4229	0.1939	0.285
171	0.5192	0.2528	0.1697	0.2588

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	4	0.6429	0.4832	4.381e-05
2	10	0.7321	0.61	6.116e-08
3	70	0.6964	0.5454	1.078e-06
4	**35**	**0.7679**	**0.6576**	**2.45e-09**
5	5	0.6964	0.549	1.078e-06
6	4	0.75	0.6286	1.281e-08
7	8	0.7143	0.5861	2.677e-07
8	20	0.6607	0.5	1.375e-05
9	3	0.5714	0.3705	0.002204
10	6	0.6607	0.5035	1.375e-05
11	30	0.6964	0.5572	1.078e-06
12	8	0.6964	0.5549	1.078e-06
13	9	0.6964	0.5611	1.078e-06
14	15	0.6786	0.5191	4.002e-06
15	5	0.6964	0.5584	1.078e-06
16	6	0.75	0.6338	1.281e-08
17	10	0.6607	0.5051	1.375e-05
18	2	0.6964	0.5611	1.078e-06
19	10	0.7143	0.5815	2.677e-07
20	20	0.7143	0.5811	2.677e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6946	0.04332	0.7679

Workflow runtime

8.498 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
```