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_GPL93).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.607	0.4154	0.1077	0.153
2	0.642	0.4699	0.1352	0.2115
3	0.6865	0.5459	0.1088	0.1578
4	0.6964	0.5565	0.08473	0.1219
5	0.6774	0.5311	0.07963	0.1125
6	0.7033	0.5683	0.09699	0.1384
7	0.6583	0.5005	0.1111	0.1648
8	0.6585	0.4961	0.0996	0.1501
9	0.6502	0.4843	0.09445	0.1413
10	0.6676	0.5105	0.08141	0.1214
15	0.6783	0.5247	0.07511	0.1153
20	0.6412	0.4735	0.07842	0.115
25	0.6677	0.5093	0.1079	0.158
30	0.6394	0.4637	0.08822	0.1377
35	0.6494	0.4789	0.09674	0.1493
40	0.6577	0.4891	0.08185	0.1298
45	0.6412	0.4587	0.1162	0.196
50	0.6245	0.44	0.08207	0.1332
60	0.6353	0.4543	0.1491	0.2351
70	0.6179	0.4219	0.1063	0.1737
80	0.6345	0.4476	0.1272	0.211
90	0.6005	0.3974	0.1226	0.1939
100	0.5888	0.3764	0.08514	0.1381
165	0.5265	0.2706	0.1244	0.1967

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	3	0.7358	0.6161	5.533e-07
2	3	0.7736	0.6733	2.513e-08
3	2	0.7736	0.6603	2.513e-08
4	6	0.7547	0.6391	1.238e-07
5	8	0.717	0.5825	2.255e-06
6	5	0.6226	0.4447	0.0007184
7	4	0.6415	0.487	0.0002658
8	8	0.7925	0.6957	4.606e-09
9	5	0.7736	0.6668	2.513e-08
10	4	0.6981	0.556	8.413e-06
11	5	0.6604	0.5023	9.108e-05
12	4	0.7547	0.6389	1.238e-07
13	9	0.7358	0.6095	5.533e-07
14	5	0.7358	0.6105	5.533e-07
15	**6**	**0.8113**	**0.7229**	**7.567e-10**
16	7	0.7358	0.6173	5.533e-07
17	6	0.6981	0.5599	8.413e-06
18	20	0.6226	0.435	0.0007184
19	10	0.6792	0.5253	2.884e-05
20	3	0.717	0.5859	2.255e-06

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.7217	0.05471	0.8113

Workflow runtime

12.008 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
```