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_GPL93).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.445	0.1914	0.1739	0.2511
2	0.563	0.3657	0.1212	0.1694
3	0.5658	0.3706	0.1602	0.2324
4	0.5658	0.3713	0.1483	0.2169
5	0.6029	0.4254	0.1629	0.236
6	0.6376	0.4753	0.1712	0.244
7	0.6565	0.4999	0.1574	0.2265
8	0.6202	0.4375	0.156	0.2202
9	0.6279	0.442	0.1459	0.2135
10	0.6468	0.4649	0.1249	0.2
15	0.6218	0.4391	0.1086	0.1645
20	0.6242	0.442	0.134	0.1992
25	0.6045	0.408	0.1527	0.2234
30	0.5712	0.3568	0.1567	0.2242
35	0.5886	0.3799	0.127	0.1874
40	0.6044	0.3995	0.1096	0.1622
45	0.5861	0.3738	0.1497	0.2188
50	0.5515	0.316	0.1357	0.207
60	0.5795	0.3639	0.1052	0.1527
70	0.5415	0.3072	0.1175	0.1774
80	0.5945	0.3838	0.1535	0.2271
90	0.5695	0.3407	0.1717	0.2609
100	0.5506	0.3069	0.09741	0.1658
165	0.4809	0.1901	0.1186	0.1979

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	7	0.6792	0.5293	2.884e-05
2	30	0.7358	0.6026	5.533e-07
3	15	0.6792	0.5228	2.884e-05
4	25	0.6226	0.4511	0.0007184
5	15	0.6604	0.4826	9.108e-05
6	9	0.7358	0.6207	5.533e-07
7	30	0.5849	0.3857	0.004204
8	7	0.6038	0.4095	0.001802
9	30	0.5849	0.3714	0.004204
10	90	0.5283	0.2744	0.03515
11	15	0.717	0.578	2.255e-06
12	45	0.6415	0.4692	0.0002658
13	20	0.6226	0.4575	0.0007184
14	20	0.6226	0.4444	0.0007184
15	8	0.6415	0.4557	0.0002658
16	20	0.6792	0.513	2.884e-05
17	25	0.6792	0.5258	2.884e-05
18	40	0.717	0.582	2.255e-06
19	10	0.6415	0.4852	0.0002658
20	**7**	**0.7358**	**0.6086**	**5.533e-07**

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6557	0.05706	0.7358

Workflow runtime

9.909 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
```