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_GPL92).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4195	0.1578	0.1603	0.254
2	0.5441	0.3289	0.1341	0.2024
3	0.5516	0.3341	0.153	0.2351
4	0.5812	0.3786	0.1752	0.2826
5	0.6053	0.4055	0.1729	0.2811
6	0.6036	0.4063	0.1776	0.2717
7	0.5753	0.3652	0.182	0.2732
8	0.6546	0.4812	0.1573	0.2395
9	0.6371	0.4601	0.1639	0.2447
10	0.6631	0.4969	0.1332	0.1986
15	0.6558	0.4876	0.1457	0.2226
20	0.6273	0.4364	0.155	0.2441
25	0.6273	0.4357	0.155	0.2448
30	0.5991	0.3926	0.1553	0.2452
35	0.5565	0.3279	0.1581	0.2487
40	0.5657	0.3406	0.1491	0.232
45	0.5899	0.3757	0.1604	0.2543
50	0.5748	0.3563	0.1475	0.2293
60	0.5292	0.2844	0.1213	0.192
70	0.5676	0.338	0.1687	0.2635
80	0.5634	0.3301	0.1188	0.1975
90	0.5563	0.3212	0.1377	0.2183
100	0.5626	0.3329	0.1146	0.1899
168	0.5326	0.2723	0.1895	0.3142

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	20	0.6364	0.4597	0.0003385
2	30	0.6545	0.4749	0.0001205
3	35	0.6727	0.4992	3.979e-05
4	6	0.6545	0.4988	0.0001205
5	3	0.5455	0.328	0.02042
6	45	0.6182	0.4092	0.0008835
7	10	0.6364	0.453	0.0003385
8	15	0.6909	0.5376	1.215e-05
9	5	0.6	0.4083	0.002147
10	5	0.6364	0.4552	0.0003385
11	7	0.6364	0.4522	0.0003385
12	20	0.5818	0.3706	0.004866
13	4	0.6364	0.4497	0.0003385
14	**10**	**0.7091**	**0.5669**	**3.424e-06**
15	7	0.6909	0.5309	1.215e-05
16	4	0.6909	0.5405	1.215e-05
17	30	0.6	0.3886	0.002147
18	5	0.6182	0.4222	0.0008835
19	7	0.6727	0.5274	3.979e-05
20	15	0.6727	0.5018	3.979e-05

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6427	0.04141	0.7091

Workflow runtime

9.544 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
```