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

Analysis of breast cancer tumor samples using 2-color cDNA microarrays (GSE5325).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.8286	0.6469	0.1756	0.3633
2	0.8143	0.6263	0.1656	0.3358
3	0.8286	0.6492	0.1313	0.2724
4	0.8286	0.6446	0.1127	0.2396
5	0.7857	0.5587	0.1684	0.342
6	0.8	0.5841	0.138	0.2872
7	0.8143	0.619	0.1513	0.3123
8	0.8429	0.6796	0.1421	0.2943
9	0.8286	0.6493	0.1475	0.3051
10	0.8143	0.6145	0.1513	0.3151
15	0.8	0.5887	0.1536	0.3135
20	0.8143	0.6135	0.1911	0.4102
25	0.8286	0.6422	0.1622	0.3447
30	0.8286	0.6447	0.1313	0.2762
35	0.8286	0.6498	0.1756	0.3586
40	0.7857	0.5585	0.1388	0.2813
45	0.7857	0.5557	0.1543	0.3175
50	0.7714	0.5098	0.1807	0.3964
60	0.8	0.589	0.1677	0.3414
70	0.7714	0.5095	0.1536	0.3433
80	0.7714	0.517	0.1807	0.384
90	0.7857	0.5371	0.1543	0.3422
100	0.7571	0.4794	0.1656	0.365
105	0.7714	0.5151	0.138	0.3065

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	3	0.8857	0.7667	6.136e-05
2	5	0.8857	0.7667	6.136e-05
3	6	0.8857	0.7705	6.136e-05
4	20	0.8286	0.6557	0.001202
5	60	0.8	0.595	0.003999
6	40	0.8857	0.7667	6.136e-05
7	15	0.7429	0.4522	0.02786
8	15	0.8857	0.7705	6.136e-05
9	**8**	**0.9143**	**0.8264**	**9.733e-06**
10	5	0.8	0.608	0.003999
11	8	0.8286	0.6441	0.001202
12	90	0.8286	0.6557	0.001202
13	2	0.7714	0.5254	0.01134
14	10	0.8571	0.7107	0.0003014
15	25	0.8857	0.7627	6.136e-05
16	40	0.8	0.595	0.003999
17	20	0.8857	0.7667	6.136e-05
18	10	0.8857	0.7742	6.136e-05
19	20	0.8857	0.7627	6.136e-05
20	1	0.8	0.595	0.003999

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.8471	0.04841	0.9143

Workflow runtime

5.259 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
```