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

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	0.6286	0.2543	0.1807	0.364
2	0.7	0.3819	0.1958	0.4199
3	0.6714	0.3136	0.1656	0.3487
4	0.7	0.3751	0.1054	0.225
5	0.7143	0.4057	0.1905	0.4067
6	0.6857	0.3394	0.1998	0.4352
7	0.6714	0.3143	0.1911	0.4018
8	0.6714	0.3262	0.1911	0.3777
9	0.6857	0.3472	0.1313	0.2708
10	0.6714	0.3162	0.1788	0.3818
15	0.6714	0.3188	0.2135	0.4406
20	0.6429	0.2607	0.2156	0.4378
25	0.6143	0.1857	0.2338	0.4927
30	0.6429	0.2515	0.2451	0.5026
35	0.5857	0.1433	0.2177	0.4276
40	0.6	0.161	0.1881	0.3681
45	0.6286	0.2069	0.1677	0.3585
50	0.6571	0.2649	0.138	0.3092
60	0.6	0.1554	0.1622	0.3328
70	0.6571	0.2621	0.138	0.3107
80	0.6714	0.2873	0.1355	0.3019
90	0.6714	0.298	0.1788	0.3861
100	0.6571	0.2541	0.1205	0.2684
105	0.5714	0.0691	0.1782	0.3826

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	3	0.7714	0.5484	0.01134
2	5	0.7143	0.375	0.06008
3	2	0.6571	0.3115	0.1974
4	3	0.6571	0.2881	0.1974
5	4	0.8286	0.6379	0.001202
6	2	0.8571	0.7107	0.0003014
7	4	0.7429	0.4706	0.02786
8	25	0.6286	0.1651	0.3067
9	20	0.7143	0.4262	0.06008
10	2	0.6571	0.3115	0.1974
11	7	0.8	0.5882	0.003999
12	15	0.7143	0.4068	0.06008
13	15	0.7429	0.4706	0.02786
14	20	0.7714	0.5172	0.01134
15	**5**	**0.8857**	**0.7667**	**6.136e-05**
16	5	0.7143	0.4068	0.06008
17	3	0.7429	0.4706	0.02786
18	2	0.8286	0.65	0.001202
19	5	0.7429	0.4615	0.02786
20	5	0.8	0.5882	0.003999

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.7486	0.07035	0.8857

Workflow runtime

4.984 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
```