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_GPL92).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4917	0.2599	0.1455	0.2233
2	0.6152	0.4309	0.1671	0.2364
3	0.6597	0.4956	0.1872	0.2755
4	0.6894	0.5492	0.1776	0.2536
5	0.7045	0.5686	0.1834	0.2653
6	0.6515	0.4875	0.1811	0.2627
7	0.6083	0.4168	0.1626	0.2375
8	0.6788	0.5208	0.1383	0.2053
9	0.6788	0.5232	0.1766	0.2563
10	0.7038	0.5586	0.1512	0.2272
15	0.6773	0.5155	0.1155	0.1823
20	0.6508	0.4722	0.1192	0.1806
25	0.6797	0.521	0.1048	0.1512
30	0.6956	0.5411	0.1262	0.1856
35	0.6789	0.513	0.09155	0.1344
40	0.6698	0.4952	0.1377	0.2054
45	0.6515	0.47	0.1161	0.1715
50	0.6076	0.4035	0.1253	0.1823
60	0.6523	0.4719	0.1208	0.1752
70	0.6674	0.4929	0.151	0.2286
80	0.615	0.4135	0.1665	0.2504
90	0.5983	0.3853	0.1434	0.2186
100	0.6432	0.4568	0.1361	0.204
168	0.6165	0.4093	0.1455	0.2202

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	**5**	**0.8**	**0.7027**	**1.599e-09**
2	9	0.7091	0.5578	3.424e-06
3	5	0.7091	0.5578	3.424e-06
4	6	0.6364	0.4481	0.0003385
5	4	0.6727	0.513	3.979e-05
6	3	0.6909	0.5394	1.215e-05
7	8	0.6909	0.5457	1.215e-05
8	25	0.7091	0.565	3.424e-06
9	10	0.7273	0.5966	8.87e-07
10	60	0.6909	0.5227	1.215e-05
11	15	0.8	0.7052	1.599e-09
12	7	0.7455	0.6163	2.106e-07
13	30	0.6909	0.5273	1.215e-05
14	8	0.7455	0.6131	2.106e-07
15	5	0.7818	0.6749	8.988e-09
16	10	0.6727	0.5254	3.979e-05
17	45	0.7636	0.6407	4.565e-08
18	9	0.7273	0.5869	8.87e-07
19	25	0.6909	0.5396	1.215e-05
20	10	0.7455	0.6282	2.106e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.72	0.04391	0.8

Workflow runtime

9.108 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
```