Feature Selection (FS) workflow report

Enrique Audain

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

Triple-Negative Breast Cancer (TNBC) proteome. Label-free deep proteome analysis of 44 (samples and technical repliques) human breast specimens.

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.68	0.5571	0.2227	0.3264
2	0.61	0.4708	0.2914	0.3758
3	0.7633	0.7	0.2471	0.3048
4	0.7917	0.7238	0.2013	0.2673
5	0.8467	0.7988	0.1798	0.2376
6	0.8667	0.8238	0.1851	0.2449
7	0.8467	0.7988	0.1798	0.2376
8	0.8667	0.8238	0.1851	0.2449
9	0.7967	0.7321	0.2007	0.2666
10	0.7967	0.7321	0.2007	0.2666
15	0.7967	0.7321	0.2007	0.2666
20	0.7967	0.7321	0.2007	0.2666
25	0.7467	0.6655	0.2066	0.2754
30	0.7717	0.6988	0.1877	0.2497
35	0.7717	0.6988	0.1877	0.2497
40	0.7717	0.6988	0.1877	0.2497
44	0.7967	0.7321	0.2007	0.2666

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	15	1	1	2.216e-07
2	6	0.8462	0.803	9.42e-05
3	7	1	1	2.216e-07
4	15	0.7692	0.7068	0.000816
5	15	1	1	2.216e-07
6	35	1	1	2.216e-07
7	15	0.9231	0.8992	6.703e-06
8	9	0.9231	0.9015	6.703e-06
9	5	0.7692	0.6977	0.000816
10	4	0.8462	0.8	9.42e-05
11	20	0.9231	0.9008	6.703e-06
12	6	0.7692	0.7023	0.000816
13	8	0.9231	0.9008	6.703e-06
14	25	0.7692	0.7068	0.000816
15	8	0.8462	0.7969	9.42e-05
16	35	0.7692	0.7023	0.000816
17	8	0.8462	0.7969	9.42e-05
18	30	0.9231	0.8992	6.703e-06
19	**6**	**1**	**1**	**2.216e-07**
20	20	0.8462	0.803	9.42e-05

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.8846	0.08824	1

Workflow runtime

2.054 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
```