Feature Selection (FS) workflow report

Enrique Audain

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

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	0.415	0.2077	0.3425	0.3246
2	0.6467	0.5464	0.3356	0.4158
3	0.58	0.4754	0.314	0.3705
4	0.6717	0.6036	0.3234	0.3572
5	0.68	0.6131	0.3451	0.3906
6	0.705	0.6321	0.3396	0.4012
7	0.7333	0.6667	0.3702	0.4444
8	0.7133	0.6417	0.3594	0.4304
9	0.7333	0.6667	0.3702	0.4444
10	0.7	0.6238	0.3583	0.4291
15	0.68	0.5988	0.3451	0.4117
20	0.655	0.5655	0.3678	0.4475
25	0.68	0.5988	0.3451	0.4117
30	0.6667	0.5917	0.3768	0.4452
35	0.6267	0.5667	0.3912	0.4098
40	0.5933	0.5167	0.4015	0.4476
44	0.4967	0.3964	0.3977	0.4485

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	8	0.8462	0.803	9.42e-05
2	6	0.8462	0.803	9.42e-05
3	**7**	**1**	**1**	**2.216e-07**
4	10	0.8462	0.803	9.42e-05
5	7	0.9231	0.9008	6.703e-06
6	25	0.7692	0.6977	0.000816
7	6	0.9231	0.9008	6.703e-06
8	7	0.9231	0.9008	6.703e-06
9	7	0.8462	0.8	9.42e-05
10	6	0.9231	0.9008	6.703e-06
11	7	0.7692	0.7023	0.000816
12	7	0.8462	0.7969	9.42e-05
13	6	0.6923	0.6	0.004876
14	15	0.9231	0.9	6.703e-06
15	8	0.8462	0.8	9.42e-05
16	30	0.5385	0.4179	0.07065
17	15	0.8462	0.7969	9.42e-05
18	4	0.9231	0.8992	6.703e-06
19	5	0.9231	0.9015	6.703e-06
20	10	1	1	2.216e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.8577	0.1067	1

Workflow runtime

1.691 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
```