Update:

1. Five-fold cross-validation wrapped logistic regression

Five-fold cross-validation combined with logistic regression were used to show the prediction performance for low-pass WGBS data. The detailed procedure is that DNA methylation data were divided into 5 equal parts and each of them as the test dataset and the remaining as the training dataset. In the training stage, prediction model was fitted with feature selection of AIC criteria with forward and backward selection in R. The detailed procedure is that we first starts with the full model and eliminates one predictor at a time, at each step considering whether AIC shows significant decrease by adding back in the variable removed at the previous step. Finally, we make the prediction with the prediction model built in training stage to test dataset and summarize the prediction sensitivity, specificity and accuracy. We found the sensitivity, specificity and accuracy in train were 65.0%, 98.7% and 94.2% and, in test dataset, they were 71.2%, 98.0% and 91.4% respectively.

Table 1. Prediction performance with Five-fold cross-validation with logistic regression with low-pass WGBS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Sen.train | Spe.train | Accu.train | Sen.test | Spe.test | Acc.test |
| 1 | 0.714285714 | 1 | 0.954545455 | 1 | 1 | 1 |
| 2 | 0.5 | 1 | 0.953488372 | 0.25 | 1 | 0.727273 |
| 3 | 0.833333333 | 1 | 0.976744186 | 0.5 | 0.888888889 | 0.818182 |
| 4 | 0.571428571 | 0.972222222 | 0.906976744 | 1 | 1 | 1 |
| 5 | 0.666666667 | 1 | 0.953488372 | 0.5 | 1 | 0.909091 |
| 6 | 0.833333333 | 1 | 0.976744186 | 0.5 | 0.888888889 | 0.818182 |
| 7 | 0.666666667 | 0.972972973 | 0.930232558 | 1 | 1 | 1 |
| 8 | 0.4 | 0.973684211 | 0.906976744 | 1 | 1 | 1 |
| 9 | 0.8 | 1 | 0.976744186 | 0.333333333 | 1 | 0.818182 |
| 10 | 0.571428571 | 0.972972973 | 0.909090909 | 1 | 1 | 1 |
| 11 | 0.833333333 | 1 | 0.976744186 | 0.5 | 0.888888889 | 0.818182 |
| 12 | 0.666666667 | 1 | 0.953488372 | 0.5 | 1 | 0.909091 |
| 13 | 0.8 | 0.973684211 | 0.953488372 | 0.666666667 | 1 | 0.909091 |
| 14 | 0.6 | 1 | 0.954545455 | 0.333333333 | 1 | 0.8 |
| 15 | 0.666666667 | 0.972972973 | 0.930232558 | 0.5 | 1 | 0.909091 |
| 16 | 0.714285714 | 0.972222222 | 0.930232558 | 1 | 1 | 1 |
| 17 | 0.5 | 0.973684211 | 0.909090909 | 0.5 | 1 | 0.9 |
| 18 | 0.714285714 | 0.972222222 | 0.930232558 | 1 | 1 | 1 |
| 19 | 0.6 | 0.973684211 | 0.930232558 | 0.666666667 | 1 | 0.909091 |
| 20 | 0.714285714 | 0.972972973 | 0.931818182 | 1 | 1 | 1 |
| 21 | 0.571428571 | 1 | 0.930232558 | 1 | 0.9 | 0.909091 |
| 22 | 0.5 | 0.972972973 | 0.906976744 | 1 | 1 | 1 |
| 23 | 0.5 | 1 | 0.931818182 | 1 | 1 | 1 |
| 24 | 0.5 | 1 | 0.953488372 | 0.5 | 1 | 0.818182 |
| 25 | 0.666666667 | 1 | 0.954545455 | 0.5 | 1 | 0.9 |
| 26 | 0.833333333 | 1 | 0.976744186 | 0.5 | 1 | 0.909091 |
| 27 | 0.571428571 | 0.972222222 | 0.906976744 | 1 | 1 | 1 |
| 28 | 0.6 | 0.973684211 | 0.930232558 | 0.666666667 | 1 | 0.909091 |
| 29 | 0.75 | 1 | 0.976744186 | 0.75 | 0.857142857 | 0.818182 |

1. Five-fold cross-validation wrapped random Forest

We applied random forest (RF) to test the prediction performance of the low-pass WGBD data and we found the overall out-of-bag prediction accuracy was 92.16%. We noticed 1 of 43 non-HCC sample was predicted to HCC while 8 HCC was predicted to non-HCC and 3 mistake classified HCC samples were all belong to early HCC. We also applied five-fold cross-validate and with 100 random resampling in RF approach and find the average sensitivity, specificity and accuracy in test dataset were 62.5%, 97.6% and 91.1% respectively.

Table: Here, show the first 40 times RF prediction details:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Spe.train | Sen.train | Accu.train | Spe.test | Sen.test | Acc.test |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 | 0.888888889 | 0.5 | 0.818181818 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0.818181818 |
| 5 | 1 | 1 | 1 | 0.833333333 | 0.5 | 0.7 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 | 0 | 0.909090909 |
| 8 | 1 | 1 | 1 | 1 | 1 | 1 |
| 9 | 1 | 1 | 1 | 0.875 | 0.5 | 0.8 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 |
| 11 | 1 | 1 | 1 | 1 | 0 | 0.818181818 |
| 12 | 1 | 1 | 1 | 1 | 1 | 1 |
| 13 | 1 | 1 | 1 | 1 | 1 | 1 |
| 14 | 1 | 1 | 1 | 0.875 | 1 | 0.9 |
| 15 | 1 | 1 | 1 | 1 | 1 | 1 |
| 16 | 1 | 1 | 1 | 1 | 0.666667 | 0.909090909 |
| 17 | 1 | 1 | 1 | 1 | 0 | 0.818181818 |
| 18 | 1 | 1 | 1 | 1 | 0.5 | 0.9 |
| 19 | 1 | 1 | 1 | 1 | 0 | 0.818181818 |
| 20 | 1 | 1 | 1 | 1 | 0.5 | 0.909090909 |

1. PCA analysis to genome-wide low-pass WGBS HCC dataset.

