Classification models for SPECT Heart data

Mingyi WEI, Jingyu Xu

Southern University of Science and Technology

2022.05.31

1. Methodology
2. Emprical Results
3. Referrence

Lgistic Output

- The results of logistic regression are shown in the figure below

```
Call:
glm(formula = train_data$V1 ~ ., family = "binomial", data = train_data,
    control = list(maxit = 100))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.043e+03  6.128e+07       0        1
V2          -4.940e-01  4.386e+05       0        1
V3          -8.978e-01  3.826e+05       0        1
V4           4.064e-01  3.295e+05       0        1
V5          -2.969e+00  3.068e+05       0        1
...          ...        ...           ...      ...
V42         -1.636e+00  2.997e+05       0        1
V43         -4.287e-01  1.855e+05       0        1
V44         -4.236e+00  7.265e+05       0        1
V45          6.585e+00  4.872e+05       0        1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1.1090e+02  on 79  degrees of freedom
Residual deviance: 2.8946e-10  on 35  degrees of freedom
AIC: 90
Number of Fisher Scoring iterations: 27
```

Figure 1: Logistic Output

Correlation analysis

- The results show that the Logistic model does not converge on the original data set. This is most likely due to the multicollinearity of the variables in the data. Therefore, descriptive analysis of the original data set is required.

Correlation analysis

- It can be seen from the correlation coefficient diagram of variables that there is a relatively large correlation between many adjacent variables, that is, the data set has obvious multicollinearity. So variable selection is required.

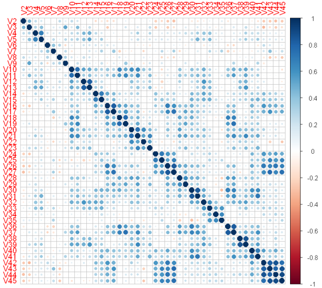

Figure 2: Correlation of Variable

LASSO

- There are many methods to select essential variables. However, it is difficult to invert a matrix $X^\top X$ as $n < p$. So we use LASSO [1] to penalize the explanatory variable coefficients by adding constraints to the loss function, which screens variables greatly.
- The LASSO estimation can be expressed as

$$\beta^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

The best tuning parameter of Lasso

- We use the lasso to achieve dimension reduction and variable selections. Figure 3 shows the path of tuning parameter so that we can get the best tuning parameter $\lambda = 0.09454452$ and 7 selected variables.

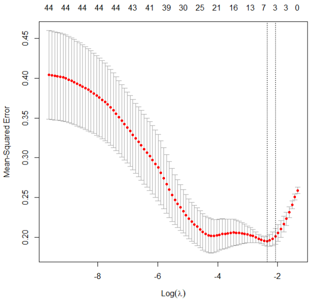

Figure 3: the best tuning parameter of lasso

Classify without dimension reduction

- We first apply different models to the original data(without dimension reduction), The prediction effect of the seven models is shown in the Table 1 below:

Table 1: Scores for seven Methods without dimension reduction

Classification Methods	AUC
SVM	0.771
KNN	0.758
Random Forest	0.778
Mlp-nn	0.688
Logistic	0.631
Naive bayes	0.804
gbm	0.775

LASSO based model: Decision Tree

- The algorithm of decision tree learning [2] is usually to recursively select the optimal feature and segment the training data according to this feature.
- The selection of this feature mainly changes through the change of maximum depth.

LASSO based model: Decision Tree

- The algorithm of decision tree learning [2] is usually to recursively select the optimal feature and segment the training data according to this feature.

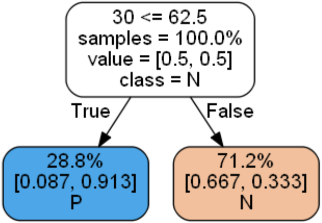

Figure 4: Decision tree, max depth=1

LASSO based model: Decision Tree

- The selection of this feature mainly changes through the change of maximum depth.

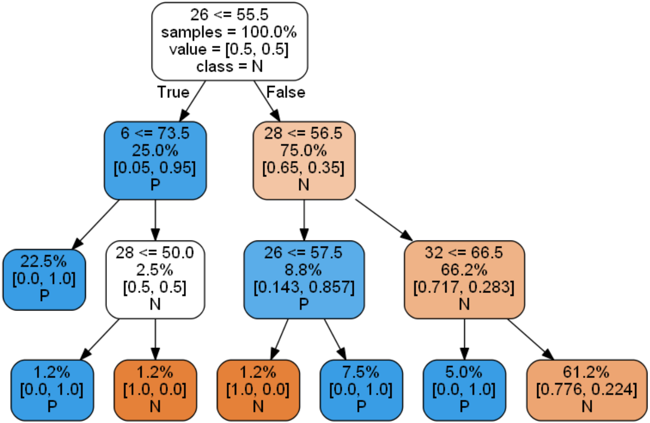

Figure 5: Decision tree, max depth=3

Model based on Lasso

- We re-run the other seven classification models, The AUC is shown in Table 2. We can compare it with Table 1.

Table 2: Results for different Classification Methods

Classification Methods	Accuracy
SVM	0.690
KNN	0.734
Naive bayes	0.810
Random Forest	0.785
Mlp-nn	0.650
gbm	0.786
Logistic	0.731

Correlation

- If prediction errors are relatively uncorrelated, we may choose Bagging to improve the accuracy; Otherwise, boosting would be a great choice.
- Thus, Checking whether they are relatively correlated is our first order of business, Figure 6 shows their relationships:

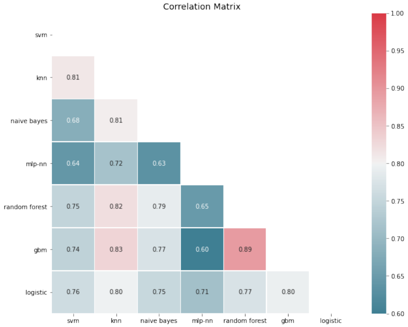

Figure 6: The correlation matrix of seven classification methods

Boosting

- We use the boosting method to train the model, and find that the ensemble performs better than most of single method, which has an AUC score: 0.803. The ROC curve is shown in Figure 7.

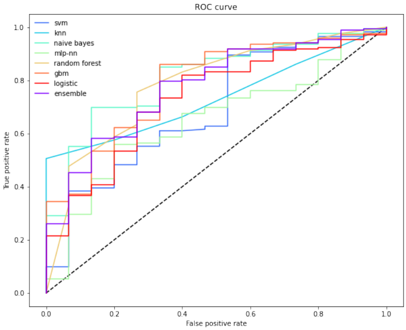

Figure 7: The ROC curve for seven classification methods and ensemble

1. Methodology
2. Emprical Results
3. Referrence

Reference

[1] Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B, Statistical Methodology. 58(1): 267–288.

[2] Podgorelec, Vili Kokol, Peter Stiglic, Bruno Rozman, Ivan. (2002). Decision Trees: An Overview and Their Use in Medicine. Journal of medical systems. 26. 445-63. 10.1023/A:1016409317640. Feature Space. Journal of the Royal Statistical Society, Ser. B, 70: 849–911. 544-549.

[3] Breiman, L.Random Forests. Machine Learning 45, 5–32 (2001).

[4] Logistic Model. In: Encyclopedia of Entomology. Springer, Dordrecht. (2004)

[5] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning20.3 (1995): 273-297.

[6] Trevor Hastie, Robert Tibshirani, Jerome Friedman(2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.), Springer, New York, 337-388

Thanks!