Coursework 1: Case Study – Feature Selection

Name: YIZE LI

ID: 2447939L

# Introduction

* This case study is mainly focus on applying proper feature engineering methods among filter, wrapper and embedded to the dataset. And then using a classifier to predict whether the subject will develop pain according to the description in this case. Task can be divided into two parts. First is to decide which method should be used. Second is to follow the usual machine learning flow, reducing dimension and train a chosen classifier using processed data.
* The dataset is preprocessed brain EEG data from SCI patients recorded while resting with eyes closed (EC) and eyes opened (EO). It can be described as 180 sets of data (in 18 groups) with 432 features (in 9 groups). The label contains two status: will or will not develop pain in the future. After applying feature engineering methods, using Leave-One-Out cross validation method to train and test on the dataset. I built a model to predict whether the subject with Spinal Cord Injury will develop Central Neuropathic Pain and got the data of accuracy, sensitivity and specificity.

# Methods

* There are mainly three ways to do feature selection: Filter method, Wrapper method and Embedded method. In this case, I will apply three methods separately and keep the rest of the code same which, in detail, contains using the same PCA function to reduce dimensions further, using the same knn classifier and LOO cross validation method to calculate accuracy sensitivity and specificity. By comparing all three feature selection methods to decide which one is better.
* First filter method measures the relevance of individual feature for dependent variable and keeps top K relevant features. K can be selected by plotting the performance of the model. Also, once K is decided, the parameter of PCA function can be decided in the same way.
* Second greedy forward method is one kind of wrapper methods. It selects features according to the baseline of the model and then updates the feature list until further action will degrade the performance.
* Third method uses L1 regularization to induce sparsity of the data, which is one way of doing embedded method. It adds a penalty against complexity to reduce the degree of overfitting or variance of a model by adding more bias. And the SelectFromModel function in sklearn can automatically decide how many features will be left.

# Results

* Filter method:

Apply SelectKBest function within the sklearn package to choose proper features. And by plotting the image of how many features we should choose, Since the accuracy is what I cared most, it can be seen from the first plot of acc that the best number is 23. And the sensitivity and specificity plots are listed below as well.

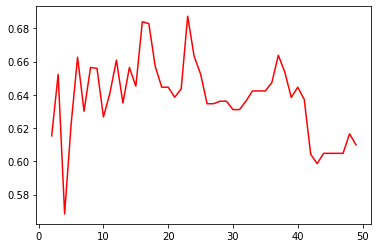


Figure 1.1 Accuracy in different occasions which the number of feature selection varies from 2 to 50

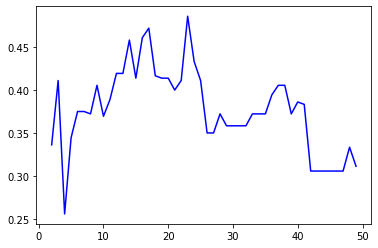


Figure 1.2 Sensitivity in different occasions which the number of feature selection varies from 2 to 50

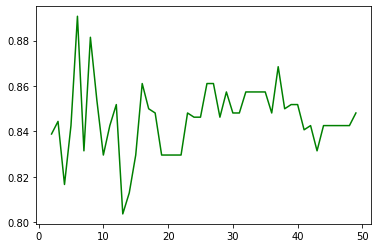


Figure 1.3 Specificity in different occasions which the number of feature selection varies from 2 to 50

And how many dimensions should be remained by PCA can be decided by the plot below. It describes the accuracy of the model with different PCA dimensions.

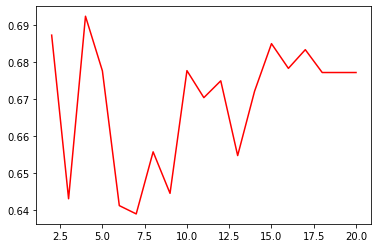


Figure 1.4 Accuracy in different occasions which the number of dimension handled by PCA method varies from 2 to 23

So the best choice is to reduce the number of dimensions to 4.

Finally, when using filter method to do feature selection work, the accuracy, sensitivity and specificity are

When using Filter Method: 0.6923120089786757 0.4277777777777778 0.9018518518518519

* Wrapper method:

Apply RFE function within sklearn package to use greedy forward feature selection. It can also be seen that when keeping 2 features the performance will be the best. And the best number of deciding how many dimensions will be left is 2(cause the data has only 2 dimensions right now).

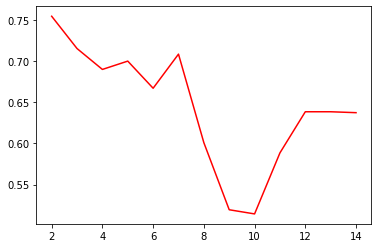


Figure 1.5 Accuracy in different occasions which the number of feature selection varies from 2 to 14

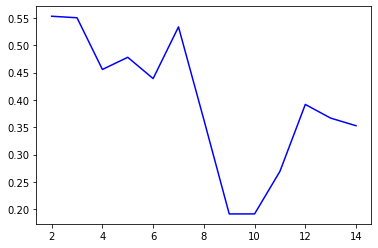


Figure 1.6 Sensitivity in different occasions which the number of feature selection varies from 2 to 14

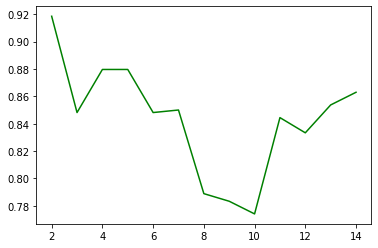


Figure 1.7 Specificity in different occasions which the number of feature selection varies from 2 to 14

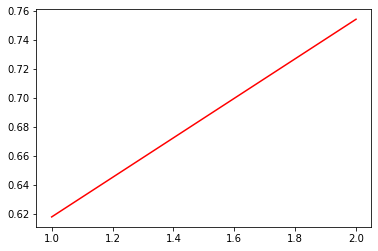


Figure 1.8 Accuracy in two occasions which the number of dimensions handled by PCA method are 1 and 2

Finally, when using filter method to do feature selection work, the accuracy, sensitivity and specificity are

When using Wrapper Method: 0.7544332210998879 0.5527777777777777 0.9185185185185186

* Embedded method:

Apply SelectFromModel function from sklearn package and use LogisticRegression to calculate the absolute value of the weights. Cause the function will automatically decide how many features will be left so the next thing I need to decide is how many dimensions should be left after the PCA function. As the figure below shows that best number is 5.

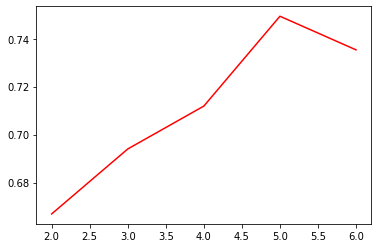


Figure 1.9 Accuracy in different occasions which the number of dimensions handled by PCA method varies from 2 to 23

Finally, when using filter method to do feature selection work, the accuracy, sensitivity and specificity are

When using Embedded Method: 0.7493827160493828 0.5416666666666666 0.914814814814815

* As all the results shown above and the table below, since the wrapper method and embedded method have very similar figures, the best feature selection choice for the dataset can be both wrapper and embedded. Besides, embedded method takes almost the same time to run as filter method, and wrapper requires much more time to run, so it is fair to say that embedded method is even a little better.

Table 1.1 Measurement of different feature selection methods

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity |
| Filter | 0.6923120089786757 | 0.4277777777777778 | 0.9018518518518519 |
| Wrapper | 0.7544332210998879 | 0.5527777777777777 | 0.9185185185185186 |
| Embedded | 0.7493827160493828 | 0.5416666666666666 | 0.914814814814815 |

# Discussion

Benefits of filter method are that it has a very low computation time and will not overfit the data. But the accuracy and other figures are not as good as other two methods. Both wrapper and embedded method can reach a better result, by taking the time cost into account, embedded method can get a relatively high accuracy and don not require so much time to run.