# 整体思路和实验结果

**目录**

[整体思路和实验结果 1](#_Toc43839255)

[一． 数据预观察阶段 3](#_Toc43839256)

[正负样本均衡性：正负样本基本均衡，不需要做处理 3](#_Toc43839257)

[特征维度大小，是否需要进行降维处理 3](#_Toc43839258)

[训练数据和测试数据的规模 4](#_Toc43839259)

[特定算法是否需要归一化 4](#_Toc43839260)

[相关性关系 4](#_Toc43839261)

[二． 单个算法模型独立运行，并分别将每个模型参数调整到最优 4](#_Toc43839262)

[随机森林算法 5](#_Toc43839263)

[深度森林算法 5](#_Toc43839264)

[支持向量机算法 6](#_Toc43839265)

[多层感知器 6](#_Toc43839266)

[逻辑回归 6](#_Toc43839267)

[三、多模型（共上述五个模型）结果融合并循环进行训练数据扩充和特征学习 7](#_Toc43839268)

[经过第一轮模型结果融合后新的训练集准确率： 7](#_Toc43839269)

[经过第二轮模型结果融合后新的训练集准确率： 8](#_Toc43839270)

[经过第三轮模型结果融合后新的训练集准确率：（准确率已经没有明现提升了） 10](#_Toc43839271)

[四、特征拆分进行协同训练 11](#_Toc43839272)

[1、取前七个特征使用随机森林进行特征学习和预测（均根据训练集的变化完成参数调优训练） 11](#_Toc43839273)

[2、取后四个特征使用随机森林进行特征学习和预测 11](#_Toc43839274)

[3、取中间8个特征进行特征学习和预测： 12](#_Toc43839275)

[4、取左右两边6个特征进行特征学习和预测： 12](#_Toc43839276)

[五、训练数据集相对测试数据集过小，使用半监督学习方式。 13](#_Toc43839277)

**单模型及模型融合准确率**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型/参数最优准确率 | 单模型最优 | 第一轮模型融合及训练数据增强 | 第二轮模型融合及训练数据增强 | 第三轮模型融合及训练数据增强 |
| 随机森林 | 袋外样本评估: 0.85  AUC: 1.000000  k折交叉验证： 0.8454 | 袋外样本评估: 0.9467  AUC: 0.9999  k折交叉验证：0.9383 | 袋外样本来评估模型: 0.9518AUC Score: 0.999918  k折交叉验证： 0.9407 | 袋外样本评估: 0.9515  AUC: 0.9998  k折交叉验证： 0.9353 |
| 深度森林 | AUC: 0.851 | AUC Score: 0.9343 | AUC Score: 0.9236 | AUC Score: 0.9255 |
| 逻辑回归 | AUC: 0.671443  k折交叉验证： 0.6304800462695199 | AUC Score: 0.9199  k折交叉验证： 0.8870 | AUC Score: 0.9272  k折交叉验证： 0.9087 | AUC Score: 0.927307  k折交叉验证： 0.9173 |
| 支持向量机 | AUC: 0.666204  k折交叉验证： 0.6259 | AUC Score: 0.920269  k折交叉验证的准确率为： 0.8915 | AUC Score: 0.927302  k折交叉验证的准确率为： 0.9117 | AUC Score: 0.927927  k折交叉验证： 0.9145 |
| 多层感知器 | AUC: 0.623450  k折交叉验证： 0.5695 | AUC Score (Train): 0.858269  k折交叉验证的准确率为： 0.7683 | AUC Score (Train): 0.898968  k折交叉验证的准确率为： 0.8233 | AUC Score (Train): 0.893741  k折交叉验证的准确率为： 0.8056 |

**多特征协同训练**

|  |  |  |  |
| --- | --- | --- | --- |
| 随机森林模型 | 袋外样本评估 | AUC | k折交叉验证 |
| 全部特征 | 0.85 | 1.000000 | 0.8454685665211982 |
| 前7个特征 | 0.819 | 1.000000 | 0.8184920342815081 |
| 后7个特征 | 0.722 | 1.000000 | 0.7114885114885114 |
| 中间8个特征 | 0.7865 | 1.000000 | 0.7859789333473545 |
| 前后6个特征 | 0.7505 | 1.000000 | 0.732958269800375 |

## 数据预观察阶段

代码：

[..\kaggle\_Classification\Data\_pre\_observation\Data\_pre\_observation.py](../kaggle_Classification/Data_pre_observation/Data_pre_observation.py)

### 正负样本均衡性：正负样本基本均衡，不需要做处理

类别输出为：

0 1099

1 901

### 特征维度大小，是否需要进行降维处理

14维特征向量，维度相对训练数据条数可以接受，不做降维处理。

### 训练数据和测试数据的规模

训练数据特征： (2000, 15)

测试数据特征： (12892, 15)

训练数据和测试数据数量差异悬殊，需要做处理，考虑三种方法：

1）多模型融合并进行融合后的训练数据扩充

2）多特征协同训练策略

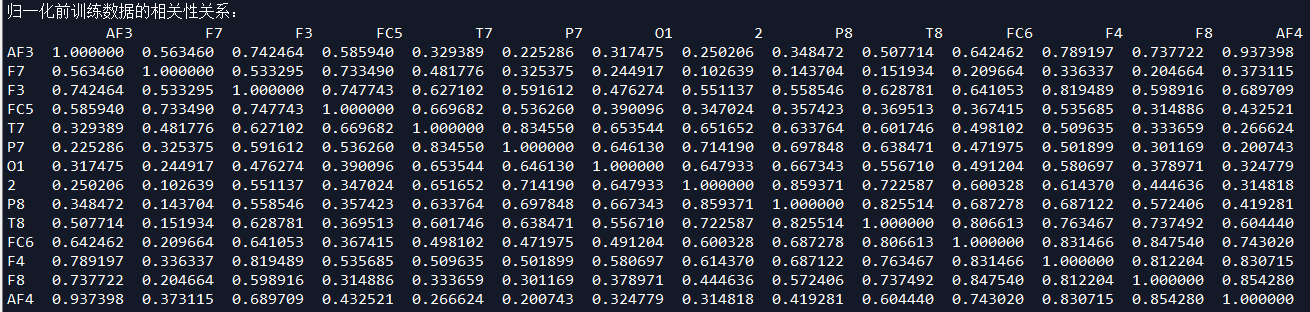
3）和无监督结合，进行半监督学习

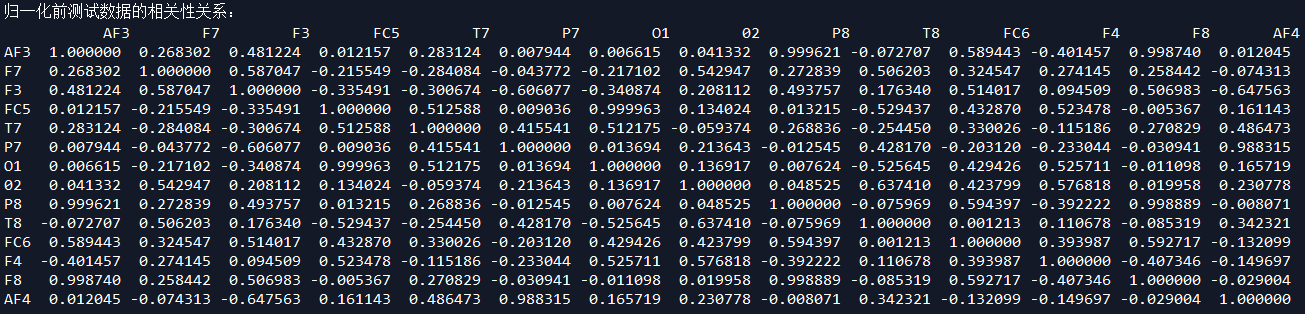
### 特定算法是否需要归一化

每个维度的特征数据级别相同，且数据大小差异小，做完归一化处理后反而不容易体现各特征的差异，效果更差，因此不进行归一化处理。

### 相关性关系

训练数据和测试数据的相关性关系不同，因此非常容易发生过拟合现象，可能在训练集结果很好，却不能较好的拟合测试集。





## 二． 单个算法模型独立运行，并分别将每个模型参数调整到最优

多参数最优组合确定：使用网格搜索方法

此时用的是原始训练和测试数据集，训练数据特征： (2000, 15)，测试数据特征： (12892, 15)

对比结果汇总到一张总表：

[..\kaggle\_Classification\single\_algorithm\_model\result.csv](../kaggle_Classification/single_algorithm_model/result.csv)

### 随机森林算法

**最优参数设置：**

alg = RandomForestClassifier(

oob\_score=True,

random\_state=10,

n\_estimators=140,

max\_depth=21,

min\_samples\_split=3,

min\_samples\_leaf=1,

max\_features=7

)

**结果准确性：**

袋外样本来评估模型: 0.85

AUC Score (Train): 1.000000

k折交叉验证的准确率为： 0.8454685665211982

**代码文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\RandomForestClassifier\RandomForestClassifier.py](../kaggle_Classification/single_algorithm_model/RandomForestClassifier/RandomForestClassifier.py)**

**结果文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\RandomForestClassifier\random\_forest\_predict.csv](../kaggle_Classification/single_algorithm_model/RandomForestClassifier/random_forest_predict.csv)**

### 深度森林算法

**最优参数设置：**

GCForest.gcForest(shape\_1X=8,window=8,n\_mgsRFtree=60,n\_cascadeRF=4,n\_cascadeRFtree=70)

**结果准确性：**

AUC Score (Train): 0.851368

**代码文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\GCForest\GCForest\_model .py](file:///E:\\data_mining\\kaggle_Classification\\single_algorithm_model\\GCForest\\GCForest_model%20.py)**

**[..\kaggle\_Classification\single\_algorithm\_model\GCForest\GCForest.py](file:///E:\\data_mining\\kaggle_Classification\\single_algorithm_model\\GCForest\\GCForest.py)**

**结果文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\GCForest\GCForest\_predict.csv](file:///E:\\data_mining\\kaggle_Classification\\single_algorithm_model\\GCForest\\GCForest_predict.csv)**

### 支持向量机算法

**最优参数设置：**

model=svm.SVC(C=5,kernel='poly',gamma=0.1,coef0=0,degree=1,probability = True)

**结果准确性：**

AUC Score (Train): 0.666204

k折交叉验证的准确率为： 0.6259705207073628

**代码文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\svm\svm.py](../kaggle_Classification/single_algorithm_model/svm/svm.py)**

**结果文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\svm\svm\_poly\_predict.csv](../kaggle_Classification/single_algorithm_model/svm/svm_poly_predict.csv)**

### 多层感知器

**最优参数设置：**

clf = MLPClassifier(activation='relu',solver='adam',alpha=1e-4,hidden\_layer\_sizes=(50,30,20),max\_iter=1200,early\_stopping=True,batch\_size=32)

**结果准确性：**

AUC Score (Train): 0.623450

k折交叉验证的准确率为： 0.5695041800304959

**代码文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\MLPClassifier\MLPClassifier.py](../kaggle_Classification/single_algorithm_model/MLPClassifier/MLPClassifier.py)**

**结果文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\MLPClassifier\MLP\_predict.csv](../kaggle_Classification/single_algorithm_model/MLPClassifier/MLP_predict.csv)**

### 逻辑回归

**最优参数设置：**

lr = LogisticRegression(penalty='l2',solver='lbfgs')

**结果准确性：**

AUC Score (Train): 0.671443

k折交叉验证的准确率为： 0.6304800462695199

**代码文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\LogisticRegression\LogisticRegression.py](../kaggle_Classification/single_algorithm_model/LogisticRegression/LogisticRegression.py)**

**结果文件：**

**[..\kaggle\_Classification\single\_algorithm\_model\LogisticRegression\LogisticRegression\_predict.csv](../kaggle_Classification/single_algorithm_model/LogisticRegression/LogisticRegression_predict.csv)**

## 三、多模型（共上述五个模型）结果融合并循环进行训练数据扩充和特征学习

该思路的流程图：（r融合时的权重基于准确率进行计算）



### 经过第一轮模型结果融合后新的训练集准确率：

训练数据特征： (7193, 15)，测试数据特征： (7699, 15)

对比结果汇总到一张总表：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\first\_round\result1.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/first_round/result1.csv)**

1. 随机森林算法

结果准确性：

袋外样本来评估模型: 0.9467537884053941

AUC Score (Train): 0.999995

k折交叉验证的准确率为： 0.9383344141066399

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\first\_round\random\_forest\_predict1.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/first_round/random_forest_predict1.csv)**

1. 深度森林算法

结果准确性：

AUC Score (Train): 0.934335

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\first\_round\GCForest\_predict1.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/first_round/GCForest_predict1.csv)**

1. 支持向量机算法

结果准确性：

AUC Score (Train): 0.920269

k折交叉验证的准确率为： 0.8915353753527722

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\first\_round\svm\_poly\_predict1.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/first_round/svm_poly_predict1.csv)**

1. 多层感知器

结果准确性：

AUC Score (Train): 0.858269

k折交叉验证的准确率为： 0.768305048705093

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\first\_round\MLP\_predict1.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/first_round/MLP_predict1.csv)**

1. 逻辑回归

结果准确性：

AUC Score (Train): 0.919912

k折交叉验证的准确率为： 0.8870914301773808

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\first\_round\LogisticRegression\_predict1.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/first_round/LogisticRegression_predict1.csv)**

### 经过第二轮模型结果融合后新的训练集准确率：

训练数据特征： (9464, 15)，测试数据特征： (5428, 15)

对比结果汇总到一张总表：result2.csv

1. 随机森林算法

结果准确性：

袋外样本来评估模型: 0.9518174133558749

AUC Score (Train): 0.999918

k折交叉验证的准确率为： 0.9407227387996617

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Second\_round\random\_forest\_predict2.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Second_round/random_forest_predict2.csv)**

1. 深度森林算法

结果准确性：

AUC Score (Train): 0.923670

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Second\_round\GCForest\_predict2.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Second_round/GCForest_predict2.csv)**

1. 支持向量机算法

结果准确性：

AUC Score (Train): 0.927302

k折交叉验证的准确率为： 0.9117709213863059

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Second\_round\svm\_poly\_predict2.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Second_round/svm_poly_predict2.csv)**

1. 多层感知器

结果准确性：

AUC Score (Train): 0.898968

k折交叉验证的准确率为： 0.8233305156382079

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Second\_round\MLP\_predict2.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Second_round/MLP_predict2.csv)**

1. 逻辑回归

结果准确性：

AUC Score (Train): 0.927201

k折交叉验证的准确率为： 0.9087066779374472

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Second\_round\LogisticRegression\_predict2.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Second_round/LogisticRegression_predict2.csv)**

### 经过第三轮模型结果融合后新的训练集准确率：（准确率已经没有明现提升了）

训练数据特征： (10508, 15)，测试数据特征： (4384, 15)

对比结果汇总到一张总表：result3.csv

1. 随机森林算法

结果准确性：

袋外样本来评估模型: 0.9515607156452227

AUC Score (Train): 0.999877

k折交叉验证的准确率为： 0.9353136770020924

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Third\_round\random\_forest\_predict3.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Third_round/random_forest_predict3.csv)**

1. 深度森林算法

结果准确性：

AUC Score (Train): 0.925586

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Third\_round\GCForest\_predict3.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Third_round/GCForest_predict3.csv)**

1. 支持向量机算法

结果准确性：

AUC Score (Train): 0.927927

k折交叉验证的准确率为： 0.9145799251791262

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Third\_round\svm\_poly\_predict3.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Third_round/svm_poly_predict3.csv)**

1. 多层感知器

结果准确性：

AUC Score (Train): 0.893741

k折交叉验证的准确率为： 0.805603322554055

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Third\_round\MLP\_predict3.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Third_round/MLP_predict3.csv)**

1. 逻辑回归

结果准确性：

AUC Score (Train): 0.927307

k折交叉验证的准确率为： 0.9173401813455077

结果文件：

**[..\kaggle\_Classification\Model\_fusion\_and\_data\_expansion\result\Third\_round\LogisticRegression\_predict3.csv](../kaggle_Classification/Model_fusion_and_data_expansion/result/Third_round/LogisticRegression_predict3.csv)**

## 四、特征拆分进行协同训练

降低模型的复杂度，减小过拟合，把特征分成多类，分别从不同角度的特征进行分类得到不同的分类器，分别用不同的分类器对测试数据进行预测。最后将这些分类器的结果进行融合，预测标签全部一致的使用加权法确定最后的类别proba。

### 1、取前七个特征使用随机森林进行特征学习和预测（均根据训练集的变化完成参数调优训练）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AF3 | F7 | F3 | FC5 | T7 | P7 | O1 |

袋外样本来评估模型: 0.819

AUC Score (Train): 1.000000

k折交叉验证的准确率为： 0.8184920342815081

训练和测试文件：

**[..\kaggle\_Classification\co\_training\data\eeg\_train1.csv](../kaggle_Classification/co_training/data/eeg_train1.csv)**

**[..\kaggle\_Classification\co\_training\data\eeg\_test1.csv](../kaggle_Classification/co_training/data/eeg_test1.csv)**

预测结果文件：

**[..\kaggle\_Classification\co\_training\result\random\_forest\_predict1.csv](../kaggle_Classification/co_training/result/random_forest_predict1.csv)**

### 2、取后四个特征使用随机森林进行特征学习和预测

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| O2 | P8 | T8 | FC6 | F4 | F8 | AF4 |

袋外样本来评估模型: 0.722

AUC Score (Train): 1.000000

k折交叉验证的准确率为： 0.7114885114885114

训练和测试文件：

**[..\kaggle\_Classification\co\_training\data\eeg\_train2.csv](../kaggle_Classification/co_training/data/eeg_train2.csv)**

**[..\kaggle\_Classification\co\_training\data\eeg\_test2.csv](../kaggle_Classification/co_training/data/eeg_test2.csv)**

预测结果文件：

**[..\kaggle\_Classification\co\_training\result\random\_forest\_predict2.csv](file:///E:\\data_mining\\kaggle_Classification\\co_training\\result\\random_forest_predict2.csv)**

### 3、取中间8个特征进行特征学习和预测：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| FC5 | T7 | P7 | O1 | 2 | P8 | T8 | FC6 |

袋外样本来评估模型: 0.7865

[0.94871795 0.22307692 0.15897436 ... 0.3 0.09487179 0.12051282]

AUC Score (Train): 1.000000

k折交叉验证的准确率为： 0.7859789333473545

训练和测试文件：

**[..\kaggle\_Classification\co\_training\data\eeg\_train3.csv](../kaggle_Classification/co_training/data/eeg_train3.csv)**

**[..\kaggle\_Classification\co\_training\data\eeg\_test3.csv](../kaggle_Classification/co_training/data/eeg_test3.csv)**

预测结果文件：

**[..\kaggle\_Classification\co\_training\result\random\_forest\_predict3.csv](../kaggle_Classification/co_training/result/random_forest_predict3.csv)**

### 4、取左右两边6个特征进行特征学习和预测：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| AF3 | F7 | F3 | F4 | F8 | AF4 |

袋外样本来评估模型: 0.7505

AUC Score (Train): 1.000000

k折交叉验证的准确率为： 0.732958269800375

训练和测试文件：

**[..\kaggle\_Classification\co\_training\data\eeg\_train4.csv](../kaggle_Classification/co_training/data/eeg_train4.csv)**

**[..\kaggle\_Classification\co\_training\data\eeg\_test4.csv](../kaggle_Classification/co_training/data/eeg_test4.csv)**

预测结果文件：

**[..\kaggle\_Classification\co\_training\result\random\_forest\_predict4.csv](../kaggle_Classification/co_training/result/random_forest_predict4.csv)**

## 五、训练数据集相对测试数据集过小，使用半监督学习方式。

先使用聚类算法进行特征学习，再利用分类算法进行标签预测。

这些特征经过kmeans聚类后效果不好：准确率只有62.98%，这样进行特征学习和数据扩充反而会使用错误标签，造成预测效果更差。

iris数据集聚类为2类FMI评价分值为：0.629817