中山大學

SUN YAT-SEN UNIVERSITY

基于学生早期行为的学业困难识别研究

模式识别期末大作业实验报告

姓 名 吴怡宁、向娅萌、杨羔

学 号 22336245、

学 院 计算机学院

专 业 计算机科学与技术

2025 年 7 月 2 日

目录

1 实验背景

1.1 问题定义描述

在现代在线教育平台中，及时识别可能存在学业困难的学生对于实现个性化干预和提高课程完成率具有重要意义。学生在学习初期的行为数据（如资源访问频率、作业完成情况等）中，往往潜藏着影响学习结果的关键信号。

本实验旨在利用学生在课程前四周内的学习行为数据，构建分类模型预测其是否存在学业困难。我们以学生最终成绩为主标签（Fail 视为“困难学生”），同时结合点击频率、活跃天数、测验提交情况和得分等多维行为特征，通过训练不同类型的分类模型，探索模型对学业风险学生的识别能力与适用性。

实验中将使用六种机器学习算法（线性分类器、非线性分类器、决策树、集成方法，聚类算法（如 K-Means、层次聚类）、神经网络）进行对比，分析其在准确性、效率、鲁棒性和可解释性等维度的表现差异，并探讨模型对不同行为特征的敏感度和预测贡献。

1.2 数据集介绍

OULAD 官网地址

1.3 算法原理介绍

1.3.1 线性分类器

线性分类器是一类通过线性决策边界将样本进行分类的模型。其基本思想是使用一个线性函数对输入特征进行加权求和，并通过阈值进行二分类。典型的线性分类器包括感知机（Perceptron）和逻辑回归（Logistic Regression）。例如，逻辑回归通过 sigmoid 函数将线性组合的结果映射到 $[0, 1]$ 区间，从而输出概率。

1.3.2 非线性分类器

非线性分类器通过引入非线性映射或核函数，将原始特征空间映射到高维空间，使得在新空间中可以使用线性分类器完成非线性分类任务。支持向量机（SVM）在使用核函数（如 RBF 核、多项式核）时就是一种非线性分类器。

1.3.3 决策树

决策树是一种树状结构的分类与回归模型。它通过对特征空间进行条件划分，将样本划分为不同的子集，最终形成一棵从根节点到叶节点的决策路径。每个内部节点表示一个特征的判定，叶节点表示分类结果。常用的划分标准包括信息增益、信息增益率和基尼指数。

1.3.4 集成方法

集成方法通过结合多个基学习器来提高模型的稳定性和预测性能。常见的集成方法包括 Bagging（如随机森林）和 Boosting（如 AdaBoost、Gradient Boosting）。随机森林通过构建多个决策树并取其多数投票结果，提升了抗过拟合能力；而 Boosting 通过迭代地训练弱分类器，并关注前一轮错误分类的样本，从而提升整体准确率。

1.3.5 聚类算法

聚类是一种无监督学习方法，用于将样本按照相似度划分为不同的簇。

K-Means K-Means 算法通过迭代优化目标函数最小化样本到簇中心的距离，来划分 K 个聚类。初始阶段随机选择 K 个中心点，接着在每轮迭代中进行样本分配与中心更新，直至收敛。

层次聚类 层次聚类通过构建一棵聚类树（dendrogram）来逐步合并或划分样本。自底向上（凝聚型）方法从每个样本开始逐步合并最近的聚类；自顶向下（分裂型）方法则从整体开始逐步分裂为子簇，直到满足停止条件。

1.3.6 神经网络

神经网络模拟生物神经元连接结构，由输入层、若干隐藏层和输出层构成。每个神经元接收来自前一层的输入，进行加权求和后通过激活函数（如 ReLU、Sigmoid）输出结果。

多层感知机（MLP） 多层感知机是前馈神经网络的典型结构，由多个全连接层构成。通过反向传播算法（Backpropagation）进行权重更新，使得损失函数最小化。

卷积神经网络（CNN） CNN 主要用于处理具有空间结构的数据（如图像），通过卷积层提取局部特征，再通过池化层降维，最终由全连接层输出分类结果。其优势在于参数共享与局部连接，适合高维输入。

循环神经网络（RNN） RNN 适用于序列数据建模。其隐层状态在时间步之间传递，能够捕捉时间上的依赖性。改进版本如 LSTM（长短期记忆网络）和 GRU（门控循环单元）能够更有效处理长期依赖问题。

2 实验流程

2.1 实验设置

本实验基于 OULAD（Open University Learning Analytics Dataset）数据集，选取课程 FFF-2013J 中的学生为研究对象，旨在利用其前 4 周学习行为预测最终是否 Fail。采用的特征包括：

- VLE 点击行为（点击次数、活跃天数、点击密度等）
- 资源使用分布（对各种类型资源的点击总数）
- 评估成绩（前 4 周测验平均分、测验次数、分数标准差）
- 注册信息（注册时间、持续天数）

2.1.1 数据预处理

- 仅保留课程代码为 `FFF`，呈现时间为 `2013J` 的数据；
- 删除特征重要性低的资源点击类特征，如 `sharedsubpage`，`dataplus`；
- 缺失值填充为 0。

2.1.2 评估标准

为了充分评估对 Fail 类学生的识别能力，本实验采用以下评估指标：

- Accuracy（准确率）
- Precision（精确率）
- Recall（召回率）
- F1-Score（综合评价）
- 特征重要性分析

2.2 线性分类器训练流程

2.3 非线性分类器（XGBoost）训练流程

在 `train_xgboost.py` 中，我们使用了 XGBoost 分类器进行建模，其核心优势在于非线性建模能力强、处理类别不平衡灵活。

模型初始化时设置了适应类别不平衡的 `scale_pos_weight` 参数，计算如下：

```
scale_pos_weight = (y_train == 0).sum() / (y_train == 1).sum()
```

模型构建如下：

```
xgb_clf = xgb.XGBClassifier(
    objective='binary:logistic',
    n_estimators=200,
    max_depth=6,
    learning_rate=0.05,
    subsample=0.8,
    colsample_bytree=0.8,
    scale_pos_weight=scale_pos_weight,
    random_state=42,
    use_label_encoder=False,
    eval_metric='logloss'
)
xgb_clf.fit(X_train, y_train)
```

预测时同样使用了阈值调整策略：

```
y_prob = xgb_clf.predict_proba(X_test)[:, 1]
y_pred = (y_prob >= 0.65).astype(int)
```

最后输出包括准确率、分类报告、AUC 分数与特征重要性排名。该模型更能捕捉复杂行为特征与测验得分的交互关系，对于提升 Recall（尤其是 Fail 类）表现较明显。

XGBoost 模型训练时间略长，但在召回率和 AUC 表现上优于随机森林，适合应用于需要高准确识别学业困难学生的场景。

2.4 随机森林模型训练流程

在 `train_random_forest.py` 中，我们首先调用了数据处理模块：

```
from data_processing import load_and_extract_features
X, y = load_and_extract_features()
```

随后，将数据划分为训练集与测试集，比例为 8:2，并采用了 `stratify=y` 保持类别分布一致。

模型使用了 Scikit-learn 的 `RandomForestClassifier`，为提高对不平衡类别的识别能力，设置 `class_weight='balanced'`：

```
rf_clf = RandomForestClassifier(
    n_estimators=150,
    max_depth=8,
    class_weight='balanced',
    random_state=42
)
rf_clf.fit(X_train, y_train)
```

预测阶段我们使用了调节阈值的方法来优化对 Fail 类的识别：

```
y_prob = rf_clf.predict_proba(X_test)[:, 1]
y_pred = (y_prob >= 0.65).astype(int)
```

最后，模型输出了准确率、分类报告和特征重要性，以便进一步分析重要行为特征。

该流程简单高效，训练时间短，适合快速迭代，并可通过特征重要性分析辅助教育干预策略设计。

2.5 集成方法

2.6 聚类算法

2.7 神经网络

3 实验结果

3.1 线性分类器

3.1.1 实验输出分析

3.1.2 特征重要性分析

3.1.3 模型局限性与改进方向

3.2 非线性分类器 (XGBoost)

3.2.1 实验输出分析

- **整体准确率为 0.856**，很好地捕捉到了早期行为与最终成绩之间的复杂关系。
- **对“非 Fail”学生（类别 0）识别性能优异**，precision = 0.90，recall = 0.92，f1-score 达到 0.91，能够准确识别大多数正常学习状态的学生。
- **对“Fail”学生（类别 1）的识别能力增强**，recall 达到 0.62，表明模型对高风险学生的捕捉能力更强。

```
Accuracy: 0.8555758683729433

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.92      0.91       434
           1       0.66      0.62      0.64       113

    accuracy                           0.86       547
   macro avg       0.78      0.77      0.77       547
weighted avg       0.85      0.86      0.85       547
```

图 1: XGBoost 模型在前 4 周数据下的分类结果

- **ROC-AUC 达到 0.873**，说明模型整体区分正负样本的能力较强，具有良好的判别性能。

```
Feature Importances:
 registration_days                0.139945
active_days_4w                    0.074135
avg_score_4w                      0.072214
resource_clicks_ouelluminate_4w   0.069077
total_clicks_4w                   0.068359
score_std_4w                      0.052600
resource_clicks_ouwiki_4w         0.052380
resource_clicks_quiz_4w           0.043555
resource_clicks_homepage_4w       0.041906
resource_clicks_resource_4w       0.040109
resource_clicks_subpage_4w        0.037549
quiz_count_4w                     0.037463
resource_clicks_page_4w           0.036478
resource_clicks_forumng_4w        0.035259
resource_clicks_oucollaborate_4w  0.035000
resource_clicks_oucontent_4w      0.034560
resource_clicks_url_4w            0.033455
click_density_4w                  0.032604
resource_clicks_glossary_4w       0.032449
date_registration                 0.030902
resource_clicks_externalquiz_4w   0.000000
```

图 2: XGBoost 模型对早期行为特征的重要性排序

3.2.2 特征重要性分析

XGBoost 的特征重要性相对更分散、稳定，避免了单一特征的“过拟合式依赖”：

- **registration_days**（0.140）：注册后在平台持续的时间仍是最关键的预测因子；
- **active_days_4w**（0.074）、**avg_score_4w**（0.072）：活跃天数和早期成绩是衡量参与度与学习表现的重要维度；
- **resource_clicks_ouelluminate_4w**（0.069）：在线课堂参与频率与 Fail 密切相关，可能反映课程互动程度；
- **total_clicks_4w**（0.068）：总体点击量是衡量投入的有效指标；

3.2.3 模型优势与改进方向

- **Fail 类识别能力较强**：召回率提升至 0.62，f1-score 提升至 0.64，更好支持对高风险学生的早期干预。
- **非线性建模能力强**：能捕捉特征之间的交互关系，适合复杂的教育行为数据；
- **整体性能更稳定**：即便在样本不均衡的背景下，也能兼顾两类样本的表现，ROC-AUC 达 0.87。
- **局限性**：
  - 部分资源特征（如 `resource_clicks_externalquiz_4w`）重要性为 0，可删除以简化模型；
  - 调参复杂，需通过网格搜索等方式进一步优化超参数；
  - 模型可解释性略差，需借助 SHAP 等工具进一步解释个体预测。

3.3 随机森林

3.3.1 实验输出分析

Accuracy: 0.8354661791590493

Classification Report]:

	precision	recall	f1-score	support
0	0.88	0.91	0.90	434
1	0.62	0.53	0.57	113
accuracy			0.84	547
macro avg	0.75	0.72	0.73	547
weighted avg	0.83	0.84	0.83	547

图 3: 随机森林模型在前 4 周数据下的分类结果

- **整体准确率达到了 0.835**，表现良好，说明模型能够较好地识别大部分学生是否会 Fail。
- **对“非 Fail”学生（类别 0）识别表现优异**，precision = 0.86，recall = 0.90，说明模型在识别正常学生方面具有很高的准确性。

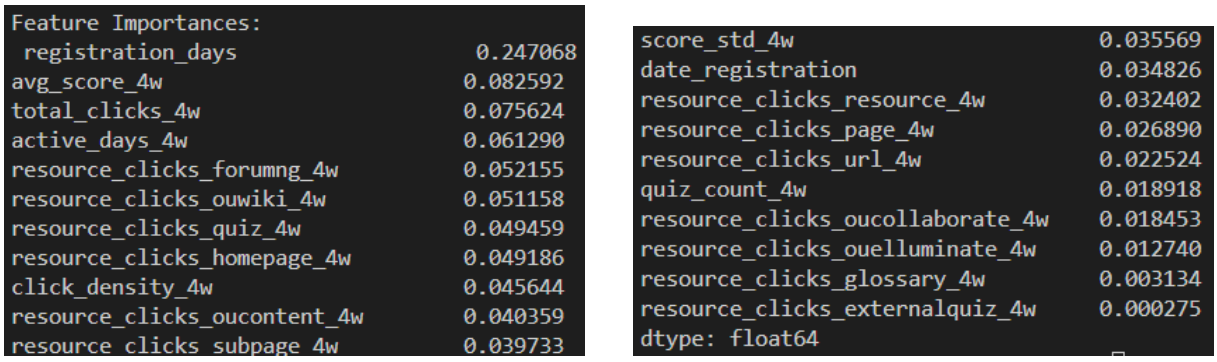

图 4: 随机森林模型对早期行为特征的重要性排序

3.3.2 特征重要性分析

随机森林相对更依赖于单一特征 **registration_days**，容易导致过拟合：

- **registration_days（0.247）**：注册后持续的天数
- **avg_score_4w（0.082）**：前 4 周平均分
- **total_clicks_4w（0.076）**：点击总数
- **activate_days_4w（0.061）**：活跃天数

3.3.3 模型局限性与改进方向

- **对“Fail”学生（类别 1）的召回率偏低**，仅为 0.53，F1 分数为 0.57，说明模型存在漏判高风险学生的风险。
- **样本不均衡影响模型表现**，由于 Fail 学生相对较少，模型更倾向于预测为多数类，导致 recall 不足。
- **特征过多但信息量有限**，一些资源点击（如 `resource_clicks_externalquiz_4w`）重要性趋近于 0，可能引入噪声。
- 对资源点击类型的依赖较强，跨课程泛化能力待验证；
- 阈值调整需根据实际业务目标精细控制；

3.4 集成方法

3.4.1 实验输出分析

3.4.2 特征重要性分析

3.4.3 模型局限性与改进方向

3.5 聚类算法

3.5.1 实验输出分析

3.5.2 特征重要性分析

3.5.3 模型局限性与改进方向

3.6 神经网络

3.6.1 实验输出分析

3.6.2 特征重要性分析

3.6.3 模型局限性与改进方向

3.7 模型优劣分析

表 1: 不同算法在多个维度下的对比分析（转置形式）

算法	准确率	问题学生 Recall	最重要特征	训练时长	数据特征敏感度	参数调整难度
线性分类器	?	?	?	快	高	低
非线性分类器	?	?	registration_days	慢	高	中
随机森林	?	?	同上	中等	中	中
集成方法	?	?		中等偏慢	**低**	**高**
聚类算法	?	?	?	中等	低	低
神经网络	?	?	?	**慢**	高	高