基于 XGBoost 特征优化与 ANN 集成的学生多维度学业表现预测

施亭妤

（福建师范大学，福建　福州　350007）

摘　要：本研究集成 XGBoost 特征选择与人工神经网络（Arificial Neural Networks, ANN），构建学生学业表现预测模型。基于多维度数据集（课堂参与、家庭背景等），通过 XGBoost 筛选关键特征，ANN 实现分类预测。实验表明，ANN 模型准确率达 0.63，优于传统算法（随机森林 0.58，SVM 0.55），为教育个性化干预提供有效工具。

关键词：学术评测成绩预测；XGBoost；人工神经网络

中图分类号：TP391.4; G449　　**文献标志码：**A

Research on Multi-Dimensional Student Academic Performance Prediction Integrating XGBoost Feature Optimization and Artificial Neural Network

SHI Tingyu

（*Fujian Normal University*, *Fuzhou* 350007, *China*）

Abstract: This study integrates XGBoost feature selection and Artificial Neural Network (ANN) to construct a prediction model for student academic performance. Based on a multi-dimensional dataset (including classroom participation, family background, etc.), key features are screened using XGBoost, and classification prediction is implemented via ANN. Experiments show that the ANN model achieves an accuracy of 0.63, outperforming traditional algorithms (Random Forest 0.58, SVM 0.55), providing an effective tool for personalized educational intervention.

Keywords: academic performance prediction; XGBoost; artificial neural network (ANN)

1　材料与方法

1.1　基准数据集

本研究使用的公开数据集来自 Github 项目 TianChi_Student_Performance（https://github.com/006Rain/TianChi_Student_Performance），包含：

（1）目标变量

Class：学生的表现类别（例如“L”低表现，“M”中等表现，“H”高表现）。

（2）核心特征（Top 5）

GradeID：学生的年级（例如“G-04”表示四年级）。

ParentschoolSatisfaction：家长对学校的满意度（“Good”表示满意，“Bad”表示不满意）。

raisedhands：学生在课堂上举手次数，反映课堂参与度。

VisITedResources：学生访问的学习资源次数，衡量课外学习的主动性。

StudentAbsenceDays：学生的缺勤情况（“Under-7”表示缺勤天数少于 7 天，“Above-7”表示多于 7 天）。

（3）数据划分

训练数据集（201 条低 / 中表现，150 条高表现），独立数据集（74 条低 / 中表现，36 条高表现）。

1.2　特征选择

本研究使用 XGboost 集成算法进行特征选择。一般来说，特征的重要性表示这个特征在构建提升树的作用。如果一个特征在所有树中作为划分属性的次数越多，那么该特征就越重要。通过每个属性分割点改进性能度量的量来计算单个决策树的重要性，并由节点负责的观察数量加权。性能度量可以是用于选择分裂点的纯度（基尼指数）或另一个更具体的误差函数。最终得到每个特征的重要性，之后可以特征排序或者进行相互比较[1]。

其相关符号表示如下：

$\tilde{y}_i^{(T)}$ 表示对第 i 个样本的预测值　　（1）

(x_i) 表示第 i 个样本的特征值　　（2）

$\sum_{j=1}^{T-1} f(x_j)$ 表示前面 T–1 棵树的累加值　　（3）

$f_j^{(T)}(x_i)$ 表示第 T 棵树的预测结果　　（4）

收稿日期：2025-04-18

作者简介：施亭妤（2005.01-），女，汉族，福建莆田人，本科，研究方向：机器学习。

1.3 K 折交叉验证

K 折交叉验证会把样本数据随机地分成 K 份（一般是均分），每次随机选择 K-1 份作为训练集，剩下的 1 份做测试集。若干轮（小于 K）之后，选择损失函数评估最优的模型和参数。其主要步骤如下：

第一步，不重复抽样将原始数据随机分为 k 份。

第二步，每一次挑选其中 1 份作为测试集，剩余 k-1 份作为训练集用于模型训练。

第三步，重复第二步 k 次，这样每个子集都有一次机会作为测试集，其余机会作为训练集。在每个训练集上训练后得到一个模型，用这个模型在相应的测试集上测试，计算并保存模型的评估指标[2]。

第四步，计算 k 组测试结果的平均值作为模型精度的估计，并作为当前 k 折交叉验证下模型的性能指标。

1.4 机器学习算法

1.4.1 随机森林

随机森林（Random Forest）乃是 Bagging（这是一种并行式的集成学习手段）的拓展形式。其基学习器被设定为决策树，众多的决策树便共同构成了森林。而所谓“随机”，主要体现在划分属性选择的随机性方面。在训练基学习器时，随机森林运用有放回采样的方式来增添样本扰动。并且，它还引入了属性扰动机制，具体而言，就是在基决策树的训练进程中，当需要挑选划分属性时，Random Forest 会率先从候选属性集合里随机抽取出一个涵盖 K 个属性的子集，随后再从这个子集中选定最为适宜的划分属性。在本次研究当中，我们借助 Python 3 的 sklearn 包来构建 RF 分类器[3]。

为了比较，我们还使用了其他四种机器学习算法，包括决策树（Decision Tree）、逻辑回归、支持向量机（SVM）、贝叶斯网络等。

1.4.2 决策树

决策树（Decision Tree）分类算法属于一种基于实例展开归纳学习的方法。其能够从所给定的无特定顺序的训练样本里，抽取出呈树状结构的分类模型。在决策树当中，每一个非叶子节点都留存着用于类别判断所运用的特定特征信息，而各个叶子节点则表征着最终判定得出的类别。从根节点至每一个叶子节点均会构建起一条用于分类的路径准则。当针对全新的样本予以测试之际，仅需起始于根节点，在每一个分支节点处施行测试操作，接着沿着对应的分支以递归形式深入子树并继续测试，持续进行直至抵达叶子节点，此时该叶子节点所指代的类别便是当前测试样本所被预测出的类别[4]。

1.4.3 逻辑回归

在逻辑回归中，我们使用对数几率函数（Logistic function）作为，对数几率函数表示形式如下：

$$y = g_{(z)} = \frac{1}{1+e^{-x}} \tag{5}$$

利用对率函数，我们可将对数几率转化为一个 (0,1) 区间内的值，除此之外，$g(z)$ 还有一个很好的导数性质：

$$g'(z) = g_{(z)}\left(1-g_{(z)}\right) \tag{6}$$

带入对率函数，得到逻辑回归表达式为：

$$y = \frac{1}{1+e^{-\left(w^T x+b\right)}} \tag{7}$$

进一步可得：

$$\ln\frac{y}{1-y} = w^T x+b \tag{8}$$

由于 y 和 $1–y$ 的和为 1，因此可将 y 和 $1–y$ 视为一对正反例的可能性，即 y 视作样本为正例的可能性，$1–y$ 为样本反例的可能性，二者比例：

$$\frac{1}{1-y} \tag{9}$$

被称为“几率”（odds），反映了样本为正例的相对可能性。对几率取对数则得到“对数几率”（logodds，亦称 logit）：

$$\ln\frac{1}{1-y} \tag{10}$$

由此看出，上式实际上是在用线性回归模型的预测结果取逼近真实标记的对数几率。因此，其对应的模型被称为“对数几率回归”（logistic regression）。

1.4.4 支持向量机

支持向量机 (SupportVectorMachine, SVM) 是一种监督学习算法，用于分类和回归，目前已广泛应用于生物信息学。其基本理念是寻找超平面，使分类区间最大化，并将数据归入不同的范畴。在这项研究中，我们使用了 Python3 的 Sklearn 包来构建 SVM 分类器[5]。

SVM 的应用非常广泛，包括图像识别、文本分类、语音识别和生物医学数据分析等多领域。由于其可靠性和准确性，SVM 在很多领域都有着广泛的应用。

1.4.5 贝叶斯网络

贝叶斯网络（Bayesian network），也被称作信念网络（belief network）或者有向无环图模型（directed acyclic graphical model），它属于一种概率图型模型。通过有向无环图（directed acyclic graphs，简称 DAGs），我们能够了解到一组随机变量及其 n 组条件概率分配（conditional probability distributions，简称 CPDs）所具备的性质。通常情况下，在贝叶斯网络的有向无环图里，节点所代表的是随机变量，这些随机变量涵盖的范围比较广，既可以是能够被直接观察到的变量，也可能是隐变量，或者是未知参数等不同类型。而连接任意两个节点的箭头有着特定的含义，它意味着这两个随机变量之间要么存在因果关系，要么属于非条件独立的情况。另外，倘若节点中的变量之间不存在箭头相互连接的状况，那么就可以说这些随机变量彼此之间是条件独立的。要是两个节点之间是通过一个单箭头连接起来的，那就表明其中一个节点为“因”，另一个节点则为“果”，并且这两个节点之间会相应地产生一个条件概率值[6]。

1.4.6 人工神经网络

人工神经网络（Artificial Neural Network，缩写为 ANN），它借助数学模型来模拟神经元的活动，是以模仿大脑神经网络的结构与功能为基础构建而成的一种信息处理体系。人工神经网络存在多层结构与单层结构的区别，在每一层之中都包含着一定数量的神经元，各个神经元之间依靠带有可变权重

的有向弧相互连接。该网络借助对已知信息开展反复的学习训练，运用逐步调整并改变神经元连接权重这种方式，进而达成处理信息以及模拟输入与输出相互关系的目标[7]。

1.5 评价指标

在本研究中，我们采用了五折和十折重交叉验证法进行模型选择和分析，如特征优化、超参数优化、学习算法比较和消融分析。此外，我们使用准确性 (ACC)、特异性 (Sp)、敏感性 (Sn) 和马修斯关系系数 (MCC) 作为评价指标，定义如下：

$$Sn = \frac{TP}{TP + FN} \tag{11}$$

$$Sp = \frac{TN}{TN + FP} \tag{12}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

2 结果和讨论

2.1 基于 XGboost 特征选择的 RF 模型

使用 XGBoost 对 17 维特征进行选择，所有特征重要性评分均较高，因此选择全部特征进行模型训练。ANN 模型在交叉验证中表现出色，AUROC 值达到 0.63，优于其他传统模型。

2.2 不同分类器建立的模型

为了验证模型的预测性能，我们采用了其他五种常用的机器学习分类器 (逻辑回归、随机森林、决策树、贝叶斯和支持向量机) 进行交叉验证，使用最优特征子集进行模型训练。对于每个分类器，我们进行五折和十折交叉验证，相应的结果如表 1 和表 2 所示，这表明基于 ANN 的模型获得了领先的成绩，具有最高的 AUROC (0.63)。

表 1 折交叉验证下各模型的性能度量对比表

Method	AUROC	AUPRC	Accuracy	F1-Score	Specificity
LR	0.60	0.58	0.57	0.58	0.56
Decision Tree	0.59	0.55	0.59	0.62	0.51
ANN	0.63	0.60	0.57	0.59	0.53
Random Forest	0.60	0.56	0.54	0.55	0.53
Naive Bayes	0.61	0.57	0.58	0.65	0.40
SVM	0.61	0.58	0.58	0.63	0.53

表 2 折交叉验证下各模型的性能度量对比表

Method	AUROC	AUPRC	Accuracy	F1-Score	Specificity
LR	0.62	0.57	0.59	0.59	0.57
Decision Tree	0.55	0.51	0.57	0.61	0.44
ANN	0.63	0.59	0.60	0.60	0.57
Random Forest	0.59	0.54	0.54	0.55	0.51
Naive Bayes	0.62	0.59	0.58	0.64	0.38
SVM	0.62	0.58	0.59	0.61	0.55

图 1 显示了基于不同算法建立的模型的 ROC 和 PRC 曲线。可以看出，ANN 的预测效果最佳。

3 结 论

在本研究中，为了预测学生的学术成绩我们提出了模型：为了对学生的学业成绩做出预测。本文在训练数据集及独立试验集上的比较显示，在本研究中人工神经网络模型

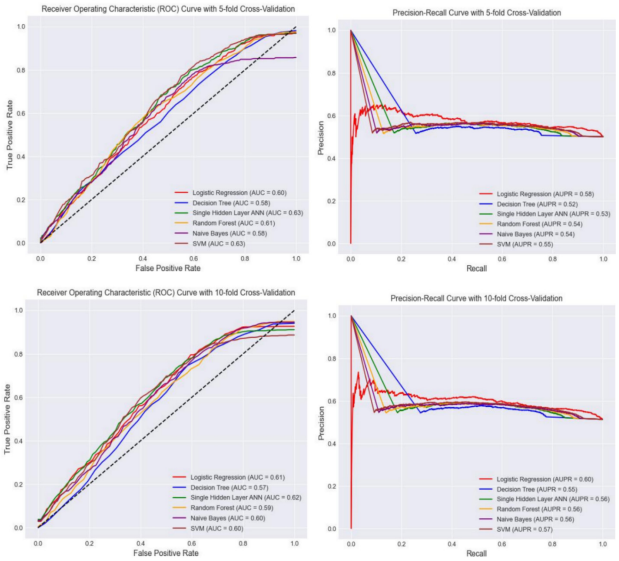

图 1 六个模型的 AUROC 曲线和 ROC 对比图

（ANN）最好在 aUROC 值为 0.63 的训练数据集中。

虽然与现有方法相比，我们的模型表现出优越的性能，但是我们展望堆叠集合算法、深度时间卷积网络或对比性学习等一些可以用来改善预测算法的其它学习方法。

参考文献：

[1] Singh H, Raghava HR, Ansari Human resource GPS(2013) enhancedmethod of antigen-primarily based on detection of the main sequences used in predicting the linear B-cell epitope. *PLoS I* 8(5):8.

[2] Li WJ Zhong L(2015) Shen WK, Zhang XF, Ying XM, Zhang W, Yan Y, Zhang F, Ying X, Zhao XM, and Zhang X F: An analysis of the amino acid anchoring pair composition of linear B-cell epitopes.

[3] Closter J M, Bjoern P, Morten N, et al. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes [J]. Nucleic Acids Research,2017,45(Web Server issue).

[4] MANAVALBAW, TANGHAVI RG, Kim MO, Lee G, Shin TH, et al. A brand-new, improved prediction method for linearly predicted B-cell epitopes. The new model was proposed by ManAVALBAJG, TONGHU, and KOH in 2014A new framework of ensemble learning.

[5] Hasan MM. Khatun MS and Kurata H(1–3), Liu MS, and Li B(5–6), are combined to create linearB-cell epitomes by combining evolutionary and sequencing characteristics in their computational identification process.

[6] Liu T. Shi K. et al. Deep learning methods increase linearB-cell epitope prediction in the area of biology and medicine (2020)10 BioData Min 13:1.

[7] EpiDope; Morris, Morris; Hayes, K.; Marz, M. et al.(2020; EpiD; EpiDope)A deep neural network used in the prediction of the linear B-cell epitope in neural networks (B cells).