心理科学进展 2024, Vol. 32, No. 11, 1814–1828
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2024.01814

·研究方法(Research Method)·

机器学习方法在测验安全领域的应用*

高旭亮 李 宁

(贵州师范大学心理学院, 贵阳 550025)

摘 要 测验安全的事后检测主要依靠统计量方法, 而新兴的机器学习方法凭借良好的检测性能与统计量方法形成互补。为了推进测验安全问题的解决, 综述了监督学习、无监督学习和半监督学习三大类机器学习方法及其衍生的集成学习、深度学习与迁移学习方法在测验安全领域的应用, 阐述了不同机器学习方法的特点, 提供了数据的获取及处理、输入特征的选择等实用的方法建议。未来研究可以从机器学习与个人拟合、机器学习与多模态数据、基于生成对抗网络的异常检测, 以及增加研究结果的可解释性几方面开展。

关键词 机器学习, 心理测验, 教育测验, 测验安全, 统计量

分类号 B841

1 引言

心理和教育测验中的作弊、加速作答或其他异常作答行为会破坏测验的可靠性以及测验分数解释的有效性(van der Linden & Guo, 2008; van Krimpen-Stoop & Meijer, 2001)。在学业成绩测验中, 会导致教师错误评估学生的学习水平, 影响教师教学(Cizek & Wollack, 2017); 在问卷调查中会导致施测者无法准确测量到目标维度, 影响问卷的信效度和对结果的解释(Arias et al., 2020; Huang et al., 2015); 这些异常行为带来的不良后果在竞争激烈的考试中更加严重, 威胁测验的安全、声誉以及施测者的筛选质量(Sinharay, 2017)。

当前根据受试者过程数据进行事后异常检测的方法主要有两种, 一种是从监控视频、眼球跟踪软件及计算机日志文件中收集的生物特征数据(Alsabhan, 2023; Ullah et al., 2019); 另一种是常规测验数据, 如被试项目反应、反应时、作答总时间等(Man et al., 2019)。前者尚处于起步阶段, 故本文重点综述常规测验数据框架下的研究。

目前研究者主要通过开发统计量方法来解决测验安全问题, 其中流行度较高的是抄袭统计量(answer copying statistics, ACS)和个人拟合统计量(person fit statistics, PFS); ACS 多是以被怀疑抄袭者和被抄袭者匹配的反应数目来构建的, 主要是对作弊中的不同类型如抄袭、项目预知等进行针对性识别(韩丹 等, 2008; Man et al., 2019)。PFS 是通过理论模型对个体的项目反应模式进行假设再与实际反应模式相对比来确定个体项目反应模式的拟合程度, 致力于对被试的不同偏差得分模式进行全面的识别(黄美薇 等, 2020; 童昊 等, 2022; 王昭 等, 2007; 张龙飞 等, 2020; 钟小缘 等, 2022; Karabatsos, 2003; Ranger et al., 2020)。胡佳琪等人(2020)和骆方等人(2020)对个体作弊和团体作弊检测方法进行了总结; 针对问卷调查中被试粗心作答的检测方法可以参考钟晓钰等人(2021)、Arthur 等人(2021)、Curran (2016)以及 Ward 和 Meade (2023)。除主流的统计量方法外, 混合模型方法也受到了一定的关注(刘玥, 刘红云, 2021)。

现如今, 普通课堂测验、计算机化自适应测验、线上交互式测验、开放性试题等多样化的考试类型伴随着不同的计分类型和评分方式不断衍生。许多测验已从传统的纸笔测验转向计算机测验, 测验过程中实时生成了大量受试者的过程数据, 这为测验安全领域带来了新的视角。统计量的研究难以支撑如此多样化的测验安全检测, 同

收稿日期: 2023-12-25
* 国家自然科学基金项目(32460212)资助。
通信作者: 高旭亮, E-mail: gaoxl9817@foxmail.com

时, 使用统计量方法纳入这些交互作用复杂的过程数据极具挑战性(Man et al., 2019), 迫切需要一种新的方法对这些非结构化过程数据进行分析。

随着数智时代的到来, 机器学习(machine learning)方法越来越多地参与到心理与教育测量研究中(刘冬予 等, 2024), 并在测验安全领域中广泛应用。机器学习算法专门用于学习数据规律并根据学习到的内容做出预测和分类(Alpaydin, 2020)。虽然机器学习存在对样本数据质量要求高等问题, 但其相比统计量方法仍有一些优势: (1)统计量方法的选择依赖于特定理论和假设, 机器学习方法的选择取决于它们检测真实数据的方式(Pan et al., 2022); (2)与受试者相关的大部分变量都可以作为输入特征训练模型, 可以充分利用过程数据; (3)模型的训练往往基于真实数据, 通过划分训练集、测试集与验证集来检验外部效度, 减轻传统方法与实证数据拟合差的问题; (4)大多数机器学习方法都具有很高的计算效率, 因此可以对大量的评估数据进行实时建模和分析。

本文根据机器学习算法的学习方式将当前测验安全领域的应用研究分为监督学习(supervised learning)、无监督学习(unsupervised learning)、半监督学习(semi-supervised learning)三大类方法进行述评, 强化学习(reinforcement learning)涉及较少, 因此未作介绍。每大类方法根据研究现状下设集成学习(ensemble learning)、深度学习(deep learning)与迁移学习(transfer learning), 我们根据各个研究使用的基础模型将其纳入不同的类别进行述评, 有些研究结合了多种机器学习方法, 我们按其使用的核心方法进行分类。第一, 我们先介绍各类机器学习方法的原理, 再对该类方法在测验安全领域的应用进行述评; 第二, 探讨了不同测验类型和异常类型下机器学习方法的适用场景, 并从已标记数据的获取、初始数据的处理、输入特征的选择等方面给出了相应的建议, 为研究者和应用者提供一定的参考和借鉴。最后对未来可研究的方向进行了展望。

2　监督学习在测验安全领域的应用

监督学习的目标是构建可以用来预测和分类的模型。在训练过程中, 模型会在已标记数据中学习从输入层(例如, 受试者的项目反应和反应时向量)到输出层(例如, 作弊反应、正常反应)的映射函数, 训练好的模型可用于预测未标记数据的输出(Alpaydin, 2020)。图 1 为监督学习示意图, X_{ij} 为被试 i 在变量 j 的反应, 输出层的“正常”与“异常”是分类标签, 监督学习是目前测验安全领域研究中最常用的方法, 适用于拥有数量和质量尚可的已标记数据的情况。我们根据当前研究将监督学习分为四部分进行介绍：(1)常规监督学习; (2)集成学习; (3)深度学习中的监督学习; (4)迁移学习。

2.1　常规监督学习

2.1.1　方法介绍

这部分主要介绍使用基本模型的研究, 监督学习中用于分类的模型主要包括：朴素贝叶斯(naive Bayes)、决策树(decision tree)、随机森林(random forest)、神经网络(neural network)、支持向量机(support vector machine)、K 近邻(K-nearest neighbors)、极端梯度提升法(extreme gradient boosting)、自适应提升法(adaptive boosting)、逻辑回归(logistic regression)、判别分析(discriminant analysis)。由于集成学习中的袋装法(bagging)和提升法(boosting)是同质分类器的集成, 是一种较为基础的模型, 因此我们将使用这些方法的研究归入了常规监督学习中进行介绍, 在集成学习板块中我们主要介绍异质分类器的集成。

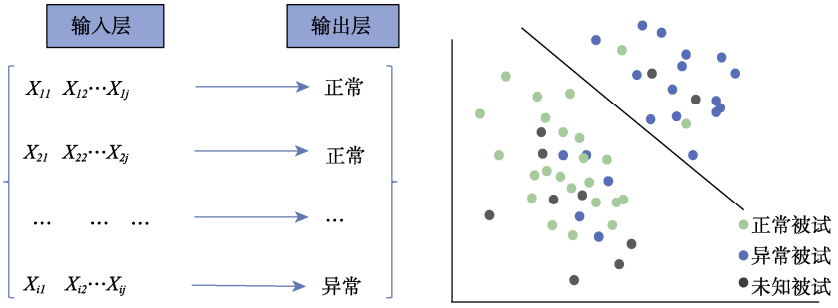

图 1　监督学习示意图

2.1.2 应用研究

Thomas (2016)、Zopluoglu (2019)以及 Man 等人(2019)使用监督学习对考试中的项目预知考生或泄露项目进行检测，这些研究都建立在已标记数据基础上。每个研究选择的输入特征不尽相同，大多都使用了模型参数值、统计量值作为输入特征，检测项目预知考生时主要关注考生的表现如项目反应和反应时，而在检测泄露项目时则需关注项目信息如项目难度、平均项目反应时间等，从而加强模型的性能。在项目预知的研究中，我们可以通过预知项目的考生来寻找泄露的项目，反之亦然，后续的研究可以从当前具有较好分类效果的研究基础上进行，如使用现有研究中标定好的泄露题目来寻找项目预知的考生。

面对没有已标记数据的情况，Zhu 等人(2022)使用了“仿真模拟”的方法进行监督学习，根据认知诊断模型(cognitive diagnostic model)模拟出正常和各种异常作答模式的考生作为训练数据，将被试的项目反应和属性掌握模式向量作为输入特征，将异常类型作为输出特征来构建模型。该方法在模拟实验中取得了优良的效果，但也有一定局限性。首先，对于异常作答的模拟只能代表“一部分”现实的情况，各种研究对于异常作答模式在现实中的表现定义也不尽相同；其次，研究在模拟异常被试时对于异常作答模式的定义十分规律，这种规律性数据就使得机器学习很容易识别作答规律性强的异常被试，但是在现实中被试表现往往更加复杂。面对无标记数据，Meng 和 Ma (2023)使用受测验条件限制较小的反应相似性指标(response similarity index, RSI)对数据中的作弊考生进行标记，再提取被标记的作弊考生特征作为机器学习输入特征来训练模型，这样模型可以对新数据中更多接近这个特征的考生进行标记，研究充分利用了统计量的优势和机器学习方法的优势，但是在选择统计量进行标记时需要衡量统计量的检验力和适用条件。这两个研究都为我们获取标记数据提供了良好的思路。

Schroeders 等人(2022)则通过实验诱发了认真与粗心的作答行为获取调查问卷中的标记数据来构建监督学习模型，但是实证研究效果较差。这说明了通过实验指导语诱发异常反应来建立模型也许是不可取的，因为很难判断现实中的参与者是否遵守了这些指示，另一方面被指示粗心回应受访者的行为方式可能与那些在外面表现出粗心回应的受访者的行为方式不同。调查问卷中的异常作答要比教育测验中的更难辨认和鉴别，因为调查问卷的项目反应并不像教育测验一样随着题目难度和受试者的能力变化，数据的规律性极差，因此难以获取高质量的异常标签，后续有学者在使用无监督学习方法得出了较好的结果，详见 3.2.2 节。

Cavalcanti 等人(2012)对开放性文本试题(主观题)抄袭进行了研究，在建立监督模型前，要对文本进行删减、规范，并将规范好的文本进行数值型转换作为输入特征。当前将机器学习文本挖掘技术应用于测验作弊的研究极少，相关领域大多研究都集中在学术剽窃。而测验中的开放性文本试题有字数少、回答零散等特点，且主要是为了检测同一考场中的考生是否存在互相抄袭行为，因此与学术剽窃检测研究的重点有所差别，学术剽窃的研究更集中于对于大型段落的再译、近义式抄袭，目的是通过机器学习识别语义特征等。建立文本抄袭检测模型之前的准备工作十分繁杂，在中文试题中我们需要进行繁琐的操作对文本进行删减和规范才能达到数值型转换的目的，而且 Cavalcanti 等人(2012)使用了特定领域的试题，专业术语的增加降低了数值转换的难度，但是这对于领域宽泛的考试却是非常致命的，各式各样的词汇增多，使得文本的数值转换工作很难进行，这也许是该领域机器学习研究较少的原因之一。类似的研究中，徐静等人(2024)使用深度学习中的卷积神经网络(convolutional neural networks, CNN)对开放式情境判断测验进行了自动化评分，分别从文档层面和句子层面对作答文本进行特征提取和分类，深度学习更适合处理自然语言，有可能成为识别文本抄袭的新途径。

2.2 集成学习

2.2.1 方法介绍

集成学习旨在结合多个基础模型的结果来开发元模型，以实现更好的预测效果(Dong et al., 2020)。集成学习主要包括袋装法、提升法和堆叠法(stacking)。袋装法和提升法都是基于同质分类器(只能使用相同的子模型开发元模型)，而堆叠法是基于异质分类器(可以使用不同的子模型开发元模型)。袋装算法是一种并行集成方法，基于每个子样本开发决策树，聚合多个决策树的结果

以找到最佳预测结果。提升法是一种按照确定性策略将弱学习算法顺序提升为强学习算法的技术(Zhou & Jiao, 2023)。

堆叠与其他两种集成学习算法的不同之处在于它整合来自不同基础模型的模型预测结果进行优化，以提高整体预测效果(Chan & Stolfo, 1997)。该算法包括两层结构，第一层中不同机器学习算法的单个基础模型分别完成训练后，第二层的元模型从第一层模型的输出中学习。图 2 为堆叠集成学习示意图。

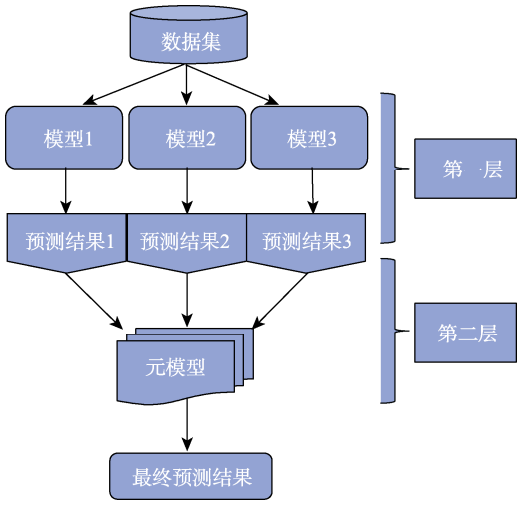

图 2　堆叠集成学习示意图

2.2.2　应用研究

Zhou 和 Jiao (2022, 2023)、Jiao 等人(2023)的研究系统的比较了集成学习模型及基础模型在不同的输入特征和重采样方式下对于项目预知考生的检测性能。研究从两方面进行了数据增强，从特征空间来看，除了常用的项目反应和反应时外，研究还在输入特征中加入了异常值检测算法得出的异常分数、统计量指标计算出的统计量值以及其他的测验过程信息并进行了特征筛选，从样本空间来看，通过 SMOTE (synthetic minority over-sampling technique; Chawla et al., 2002)达到训练样本类平衡(例如，训练数据中有 100 个被试数据，仅有 5 人异常，使用这样的样本训练模型会导致预测结果出现偏差)。结果显示，堆叠、类平衡和包括增强数据的模型效果更好。Pan 和 Wollack (2023)、Pan 等人(2022)利用集成学习的思路，使用不同的数据子集训练模型最后合并训练结果。Zhen 和 Zhu (2024)则将表现最优异的基础模型进行集成来达到最佳效果。

项目预知与题目泄露研究中模型的预测效果随着研究的复杂程度不断增加，从结构较为简单的单一监督模型，逐渐衍化成多个基础模型做比较，再到使用集成学习以及开发堆叠、混合集成学习元模型；模型输入特征从单一的受试者项目反应和反应时到各种数据增强特征的对比。可以发现，集成算法、类平衡、数据增强以及特征筛选都给机器学习模型性能带来了不小的提升，如果说 2.1 节中的基础模型是“普通电脑”，那么有了这些方法加持的研究可谓是“高性能工作站”。在进行研究时可以充分利用这些方法提升模型的性能，但是需要注意，过于复杂的方法和过多的输入特征需要极高的算力支持。

2.3　深度学习中的监督学习

2.3.1　方法介绍

深度学习是机器学习的一个分支，通过训练多层次的神经网络来自动学习输入数据的特征，并根据这些特征进行预测或分类。深度学习领域涵盖了监督学习、无监督学习和半监督学习算法：其中，深度神经网络(deep neural networks, DNN)、卷积神经网络和循环神经网络(recurrent neural networks, RNN)属于监督学习的代表性算法，深度神经网络擅长处理结构化数据，卷积神经网络专门处理图像数据，而循环神经网络擅长处理序列数据(Goodfellow et al., 2016)，深度学习在测验安全领域中已有诸多应用。

2.3.2　应用研究

Zhen 和 Zhu (2024)比较了 12 种基础模型对于项目预知考生检测的性能，他们发现深度神经网络模型 TabNet 效果优于其他基础模型，而且该模型无需超参数调整，该模型与基础模型中表现同样良好的 AdaBoost 模型集成后的 TabNet-AdaBoost 模型还超越了其他研究中同一批数据的集成学习模型性能(Zhou & Jiao, 2023)。在没有堆叠集成学习和增强数据的情况下，深度神经网络也许是处理监督分类任务的良策。

深度学习中的长短期记忆网络(long short-term memory, LSTM)是循环神经网络的一种变体，非常适合处理带有时间标签的序列数据，它可以学习数据的内在模式和结构，对未来值进行预测。研究者往往通过比较未来值与实际值的差异来判

断受试者的异常作答行为，比较典型的研究有：Tiong 和 Lee (2021)使用 LSTM 分析考生的得分和反应时，观察其是否出现超高正确率的快速作答，一旦出现异常则会给考生重新分配题目进行作答；Kamalov 等人(2021)使用 LSTM 根据考生平常的测验和期中考试的成绩来预测期末考试的成绩，然后应用异常值检测算法来识别实际与预测成绩之间的异常，时间序列数据的预测示意见表 1；Tang 等人(2023)则使用点击流(应用程序中用户操作的精确日志)建立预测模型，从而根据预测结果寻找“非典型受试者”；Alsabhan (2023)则做了更综合的研究，结合了考生操作日志、不同时期的考试成绩等序列数据识别作弊受试者。从研究中来看，时间序列数据既可以是一场考试中考生做每道题的数据，也可以是间隔一段时间后的整体考试成绩，还可以是计算机上的操作日志数据，我们可以记录考生的时间序列行为和成绩数据从而对其未来值进行预测，再从考生下一步的实际行动判断其是否符合未来值，一旦考生严重偏离典型模式，就可能是出现了异常作答反应。

表 1 时间序列数据预测示意

ID	时间点 1 成绩	时间点 2 成绩	时间点 3 成绩	时间点 4 成绩
1	72	76	79	?
2	81	75	71	?
3	90	91	93	?

注：此表仅为示意，不代表研究实际内容。

2.4 迁移学习

2.4.1 方法介绍

迁移学习可以应用于绝大部分数据样本甚至所有样本都无标记的情况，这时可以寻找一些相似的有标记数据进行迁移学习。迁移学习可以将知识从一个领域迁移到另一个领域中，其基本思想是利用已学习的知识(通过在源领域进行学习得到的模型)来帮助改善在目标领域上的学习任务，即使它们的输入空间或输出空间有所不同(项目长度、内容)，图 3 为迁移学习示意图，在测验安全领域中模型效果成功迁移的关键在于训练数据集和目标数据集中的作弊流行率和数据分布的相似性，迁移学习的详细内容可以参考 Weiss 等人(2016)。

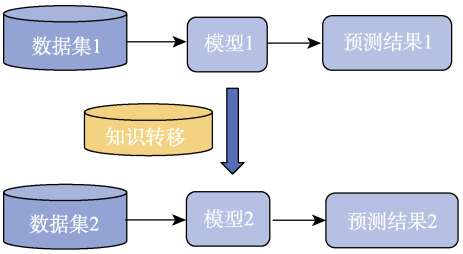

图 3 迁移学习示意图

2.4.2 应用研究

Ranger 等人(2023)将整套题目拆分为测验长度相近的三对数据集作为训练集与目标集，以此比较迁移学习的转移效果。首先使用多元双样本检验对所有数据集对(训练数据集与目标数据集)的指标联合分布相似性进行检验，然后使用训练数据对模型进行训练。可以将训练好的模型直接应用到目标数据集，也可以通过半监督学习的自训练算法(self-training)使训练好的模型不断适应目标数据集，将目标数据集中异常信号最强烈的被试不断纳入模型进行训练，重复步骤直至模型将目标数据集中的数据完全标记。

该研究为相似数据集之间的模型训练效果迁移提供了良好的借鉴，虽然研究中的训练集与目标集的统计量分布并不完全相同，但是转移后的效果仍然比使用无监督方法好得多。在拥有与目标数据集相似性较高的标记数据集时，可以尝试使用迁移学习，但是要尽量确保作弊的流行率处在同一水平，项目的长度和内容也要尽量一致，这样我们可以直接迁移学习好的模型参数到目标数据集，减少特征选择和特征变换的工作量。如果数据集之间相差较大可能要对源模型进行一定的调整来适应对目标任务的需求，这时迁移学习的工作量就会比较大，模型的迁移效果难以得到保证(Weiss et al., 2016)。对于测验安全领域来说，迁移学习的要求比较苛刻，使用时要注意数据集之间的相似性。

3 无监督学习在测验安全领域的应用

无监督机器学习评估不同变量之间的相似性，以寻找嵌入数据中的潜在结构或集群。通过评估输入变量(例如，过程数据)以揭示数据中的潜在模式，将相似的考生归入同质群组，或将空间划分为高密度(常规数据)和低密度(异常值)区域

(Alpaydin, 2020)。尽管无监督学习不会明确搜索出作弊者, 但它会将作弊者构成单独的集群。图 4 为无监督学习示意图, 图左为去掉输出层的图 1。根据当前研究, 在无监督学习中我们主要分为两类进行介绍：(1)常规无监督学习; (2)深度无监督学习。

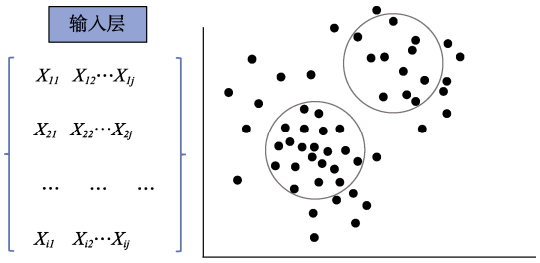

图 4　无监督学习示意图

3.1　常规无监督学习

3.1.1　方法介绍

当前研究主要使用无监督学习中的聚类(clustering)和异常值检测(anomaly detection)。聚类算法旨在将数据点分组成具有相似特征的簇, 常见的聚类算法包括 K 均值聚类、层次聚类、密度聚类等; 异常值检测算法用于识别数据中与大多数数据显著不同(模式或分布)的观测值, 这些观测值被称为异常或离群点, 可以计算一个异常值分数并设定一个阈值来判断哪些数据点为异常值, 常见的异常检测算法包括孤立森林(isolation forest)、马氏距离(mahalanobis distance)等, 详细的异常值检测方法可参考 Hodge 和 Austin (2004)、Gorgun 和 Bulut (2022)以及 Zimek 等人(2012)。

3.1.2　应用研究

Kim 等人(2016)使用已标记数据集中的项目预知考生来进行购物篮分析(无监督学习中的关联规则挖掘方法, 旨在发现特征变量之间的关系), 可以有效的分析作弊考生的共同背景特征, 这样有利于分析这些特征与被标记的被试密切相关的原因; Liao 等人(2021)使用 *K* 均值算法识别不同作答模式的考生并将其聚类, 再根据过程数据进行具体的分类; Gorgun 和 Bulut (2022)使用异常值检测方法检测了智能交互式的个性化学习系统中的异常受试者; Man 等人(2019)使用聚类方法检测项目预知考生; Man 等人(2019)与 Pan 和 Wollack (2021) (以下称 PW21)都使用了聚类方法检测了同一个数据集(Cizek & Wollack, 2017), 前者将考生的各种过程信息直接作为输入特征进行聚类, 而后者则先根据作答正误和反应时将反应极快且回答正确的数据点标记为异常(异常矩阵见示意图 5), 再根据考生作答的相似性将项目预知考生聚类, 再从聚类后的考生组作答的异常模式中识别泄露项目。

作答正误

	I1	I2	I3
E1	×	√	√
E2	√	√	×
E3	√	√	×
E4	√	√	√

+

反应时间

	I1	I2	I3
E1	慢	快	快
E2	快	快	快
E3	快	快	慢
E4	慢	慢	快

=

异常矩阵

	I1	I2	I3
E1	0	1	1
E2	1	1	0
E3	1	1	0
E4	0	0	1

图 5　异常矩阵的构建

注：该图仅作思路的示例。异常矩阵中的 0(1)代表正常(异常), 快慢是根据该考生在所有题目的平均作答时长和所有考生在此题目上的平均作答时长来定义的, 如果考生 E2 在 I1 题目的作答速度远超平均水平且答对(具体定义见 PW21), 则会被标记为 1 (异常), 生成异常矩阵后根据反应的相似性, 考生 E2 和 E3, 题目 I1 和 I2 (图中加粗)则会被聚为同一类。

Pan 和 Wollack (2023) (以下称 PW23)又在此研究基础上进行了改进, 同时对项目预知考生和泄露项目进行聚类。研究使用集成学习思路, 对不同的数据子集重复聚类和标记的过程后将标记结果合并, 最后使用自编码器(autoencoder)提供一个信度值, 从而计算考生实际作答情况与预期重构的情况之间的差距。Pan 等人(2022)将项目预知考生与题目泄露的检测拓展到了计算机化自适应测验当中, 结合了上述研究的特点, 见 4.2 节。

在测验安全领域的研究中不论是监督还是无监督学习使用的输入特征都是比较相似的。从一系列研究中可以发现, 无监督学习在相当程度上起到辅助和初步识别的作用。可以使用聚类方法对考生类别进行划分, 也可以尝试使用异常值检测方法找到异常被试, 两种方法都可以作为对异常受试者的初步识别, 而后再根据过程数据手动分类; 也可以先根据异常被试的明显异常特征手动分类再使用无监督学习进行进一步分类。无监督学习在测验安全领域的应用非常广泛。

3.2　深度学习中的无监督学习

3.2.1　方法介绍

自编码器是当前研究中使用较多的深度无监

督学习方法，作为一种异常检测方法，自编码器旨在滤除数据中的噪声。自编码器往往被用来学习数据中的模式，例如正常被试的作答反应模式，它将观察到的作答反应压缩为潜在的低维表示，并通过学习到的模式对原始数据进行重建，训练好的自编码器同样可以对新数据进行重建(Goodfellow et al., 2016)。重建良好的作答反应表明自编码器已成功学习了这些反应的内部结构。相反，无法很好重建则表明此类作答反应本质上与自编码器学习的底层结构不同。因此自编码器非常适合于检测教育测验或者是问卷调查中的异常反应，因为异常反应的内部结构往往都比较混乱。

3.2.2 应用研究

Welz 和 Alfons (2023)提出了一种基于对异常可能表现出来的三个维度(不一致性、不变性、快速反应)的新方法来识别每个参与者开始粗心应答的节点，将三个维度的证据整合在一起，从而为每个项目构建出一个基于变化点的检测分数，用以衡量某个受访者是否已开始对该项目做出异常应答并将每个应答者的应答划分为准确应答段和异常应答段[与变点分析(张龙飞 等, 2020)有相似之处]。该研究使用自编码器对观察到的反应进行重建来测量内部一致性维度，如果被试粗心作答，则算法就会出现较差的学习情况。通过计算每个被试的重建误差来比较重建与实际的差异，较大的重建误差代表了被试出现异常作答；使用回答变异性来测量第二个维度，一旦受访者开始通过直线或模式作答行为时，就会出现变化点；第三个维度反应时间测量的是受访者花在调查表每一页上的时间或花在每个项目上的时间。在实证研究中该方法取得了理想的效果，这也一定程度上证明了使用自编码器检测问卷粗心作答这种结构较为混乱的数据是有很大潜力的。Pan 等人(2022)以及 PW23 的研究中都使用了自编码器来重建了实际作答与预期作答的差距，从而提供方法的置信度。

4 半监督学习在测验安全领域的应用

4.1 方法介绍

当数据集中只有一小部分标记样本但有大量未标记样本时，半监督学习是一种很有效的方法。半监督学习通过将未标记数据纳入模型训练来更好地捕获数据分布的特征，从而提高模型的性能(Zhu & Goldberg, 2009)，见图 6。其中，自训练算法是半监督学习中最常见的方法，也是当前研究中使用最多的方法，它使用初始标记数据训练模型，然后用模型预测未标记数据，并将高置信度的预测结果作为新的标记数据添加到训练集中，这个过程重复进行，直到完全标记所有数据或达到停止条件。

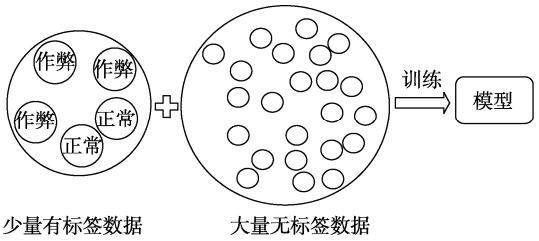

图 6 半监督学习示意图

4.2 应用研究

相当一部分研究为机器学习方法在项目预知检测上的前景提供了有力的证据，但现有的方法多为纸笔测试而设计，PW21 与 PW23 使用无监督聚类方法检测项目预知或题目泄露，Pan 等人(2022)则成功将机器学习的方法应用到更容易受到项目预知影响的计算机化自适应测验(Computerized Adaptive Testing)中。计算机化自适应测验是一种量体裁衣的测验形式，对于每一个受试者所出的题目内容和数量都不同，而且题库推荐系统是根据当前题目的难易以及受试者能否正确作答来出下一题目，如果被试在较难的题目拥有项目预知能力，那么系统就会推荐更难的题目，因此得出的测验结论将会毫无价值。

同 PW23 一致，Pan 等人(2022)先根据作答正误和反应时将一组项目反应标记为正常和异常，标记好数据后通过自训练算法迭代训练支持向量机分类器：输入特征使用考生水平的中心作答对数响应时间与项目水平的中心作答对数响应时间，用训练好的分类器对无标签的被试数据进行分类，不断从分类好的被试中挑选极端的异常被试样本添加到训练数据集中继续迭代训练分类器，直到剩余未标记数据之间的反应时间方差小于没有预知能力被试数据的预期反应时间方差时，才停止分类过程。为了防止不同数据集上的检测性能差异，采用集成学习的思路，对多个数据子集分别

标记后合并为最终的检测结果, 最后通过自编码器算法提供了一个信度值(同 PW23)。在 Ranger 等人(2023)的迁移学习研究中也用到了自训练算法, 研究者将其应用到模型效果的迁移中, 通过自训练算法不断适应目标数据集。

5　三种方法在测验安全领域应用的综合分析与使用建议

本节主要根据当前测验安全领域的机器学习研究对三大类方法进行了总结, 并给出了一些使用建议: 不同测验类型和异常类型下方法的选用、初步的数据处理、输入特征的选择以及如何获取已标记数据, 供测验安全领域研究者进行参考。由于篇幅限制, 在综述每种方法的应用时并未在文中详尽介绍研究具体使用的机器学习方法, 在表 2 中我们对其进行了总结。图 7 为方法选用流程, 以便对照第 5 节具体内容进行参考。

5.1　三种方法的综合分析

机器学习方法整体上最大的局限性是受数据的数量和质量影响, 数据是模型的营养, 如果数据质量低下或数量较少, 任何方法的效果都不会太好。监督学习方法作为本领域中最常用的一种方法, 可以使用已知的异常和正常样本进行训练, 从而建立一个分类模型来识别新数据中的异常行为。监督学习最大的优势在于一旦有质量和数量尚可的标记数据就可以实现很好的预测效果, 但这一切都要建立在有充足标记数据的情况下, 而测验安全领域由于道德、法律更因为心理过程的潜在性, 很难得到高质量的异常标记数据。监督学习的优势还在于除了基础模型外还可以使用集成学习加强模型性能, 但是集成学习需要繁复的调试超参数, 需要较好的算力才能支持大规模高维数据。使用深度学习中的长短期记忆网络进行时间序列数据的预测则需要数据中的每个数据点都带有时间标签。

无监督学习方法可以在没有已标记异常样本的情况下, 学习数据中的模式和结构, 并识别出偏离典型的异常作答反应模式。无监督学习的局限性在于无法明确量化结果, 但是对我们初步分类受试者以及了解被试的潜在作答模式或背景特

表 2　机器学习在测验安全领域应用的具体方法及文献

方法类型	具体方法	测验及异常类型
监督学习	决策树(Cavalcanti et al., 2012); 神经网络(Zhu et al., 2022); 梯度提升法(Schroeders et al., 2022); 二次判别分析(Ranger et al., 2023); 支持向量机(Thomas, 2016; Pan et al., 2022); 极端梯度提升(Zopluoglu, 2019); 支持向量机、K 近邻、随机森林(Man et al., 2019); 支持向量机、决策树、逻辑回归、朴素贝叶斯、判别分析、神经网络、梯度提升、随机森林模型构成的堆叠及混合集成学习(Jiao et al., 2023; Zhou & Jiao, 2022, 2023); 决策树、逻辑回归、朴素贝叶斯、二次判别分析、神经网络、梯度提升、随机森林、K 近邻、多层感知机、自适应提升、高斯过程、深度神经网络 TabNet (Zhen & Zhu, 2024); 长短期记忆网络(Alsabhan, 2023; Kamalov et al., 2021; Tang et al., 2023; Tiong & Lee, 2021); 逻辑回归、线性判别分析、二次判别分析、K 近邻、朴素贝叶斯、支持向量机、决策树、随机森林、自适应提升和神经网络(Meng & Ma, 2023)	教育测验作弊: (Alsabhan, 2023; Cavalcanti et al., 2012; Jiao et al., 2023; Kamalov et al., 2021; Man et al., 2019; Meng & Ma, 2023; Pan et al., 2022; Ranger et al., 2023; Tang et al., 2023; Thomas, 2016; Tiong & Lee, 2021; Zhen & Zhu, 2024; Zhou & Jiao, 2022, 2023; Zopluoglu, 2019) 教育测验作弊、随机作答、睡眠效应: (Zhu et al., 2022) 调查问卷粗心作答: (Schroeders et al., 2022)
无监督学习	层次聚类(Pan & Wollack, 2021; Pan & Wollack, 2023); K 均值聚类(Liao et al., 2021); K 均值聚类、高斯混合模型、自组织映射聚类(Man et al., 2019); 独立森林、椭圆包络、单类支持向量机、密度聚类(Jiao et al., 2023; Zhou & Jiao, 2022); 高斯混合模型(Ranger et al., 2023); 高斯混合模型、贝叶斯高斯混合模型、独立森林、马式距离、局部异常值因子和椭圆包络(Gorgun & Bulut, 2022);核密度估计(Kamalov et al., 2021); 自编码器(Pan et al., 2022; Pan & Wollack, 2023; welz & Alfons, 2023);购物篮分析(Kim et al., 2016)	教育测验作弊: (Gorgun & Bulut, 2022; Jiao et al., 2023; Kamalov et al., 2021; Kim et al., 2016; Man et al., 2019; Liao et al., 2021; Pan & Wollack, 2021; Pan & Wollack, 2023; Pan et al., 2022; Pan & Wollack, 2023; Zhou & Jiao, 2022) 调查问卷粗心作答: (welz & Alfons, 2023)
半监督学习	自训练算法(Pan et al., 2022; Ranger et al., 2023)	教育测验作弊: (Pan et al., 2022; Ranger et al., 2023)

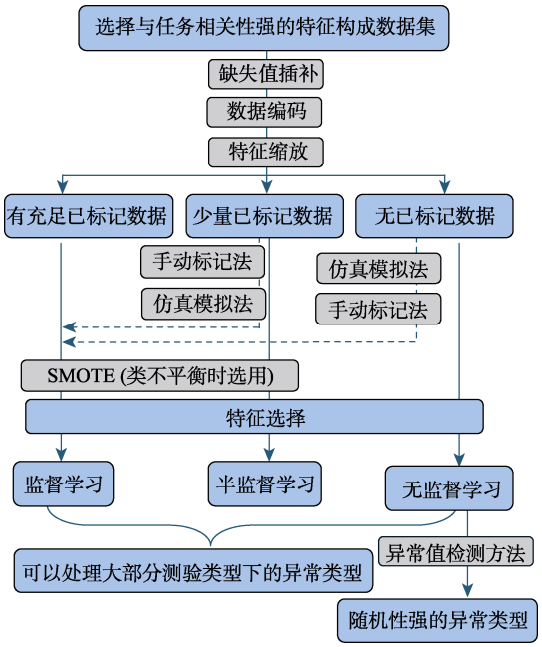

图 7 方法选用流程图

征起到很大作用。

半监督学习中的自训练算法可以根据部分已标记数据对剩余无标记数据进行预测，而这种情况在测验安全领域十分常见，因此值得引起重视。在模型学习过程中，用于训练模型的这一部分已标记数据的质量是至关重要的，因为它影响着模型的学习效果，影响着模型对于剩余大部分无标记数据的识别，因此选择已标记数据一定要非常谨慎。

总之，各种机器学习方法各有优劣，使用需视情况而定，可以将不同的机器学习方法相结合以达到更好的检测效果。

5.2 三种方法在不同场景的选用

当前的研究中的异常类型总共可以分为以下两种：(1)教育测验中的异常反应，如作弊、随机作答、睡眠效应等；(2)调查问卷测验中的粗心作答(受试者由于动机低下而随机作答、直线作答或规律作答等)。在基于计算机的测验中我们多数时候能获得的数据都包含最基础的项目反应、反应时，部分测验还会提供诸如考生的修改答案次数、情绪、点击流等更丰富的过程数据，我们往往是根据数据的情况选用不同方法(有无已标记数据？已标记数据的质量?)。由于机器学习是学习数据规律的方法，因此这三种方法在多数的测验类型和异常类型下都是适用的，只是我们选择的输入特征会有一些不同的侧重点。例如，如果我们想检测项目预知，我们可以重点关注考生快速答对的项目反应时并作为输入特征；如果我们想检测被试在测验尾部的疲劳作答，我们可以侧重于将测验尾部的项目反应等变量作为输入特征。

有一部分异常类型有很强的随机性，教育测验中的受试者在作答动机不强时可能会对任意题目进行随机作答，这导致项目反应和反应时等过程数据非常混乱；另一种是调查问卷中的粗心作答，调查问卷与教育测验的过程数据有着显著的不同，在教育测验中考生在项目上的项目反应和反应时遵循着随着题目难度上升则分数降低、反应时增加的基本规律，而在调查问卷中除了明显异常的连续一致作答和规律作答，我们很难去判断其是否认真作答。因此面对这些随机性强、过程数据无明显规律的异常类型，常用的监督学习对其效果并不明显。目前比较有效的方法是无监督学习中的异常值检测方法，尤其是深度无监督学习中的自编码器在当前研究取得了较好的效果(Welz & Alfons, 2023)。在异常检测中，自编码器通常被用来学习正常数据的表示，训练完成后自编码器可以用来重建新的输入数据。如果重建误差(即重建的数据与原始数据之间的差异)超过了某个阈值，就可以将该输入数据标记为异常，而随机性强的异常反应往往结构十分混乱、重建误差较大，因此可以较好的识别出来。

5.3 如何获取已标记数据

在面对异常受试者检测时，无论是使用半监督或者监督学习，我们往往会面临真实数据中仅有小部分或者完全没有已标记样本的难题，我们通过对现有研究的总结提出了三点方法来获取已标记数据，以便研究者参考。

(1)仿真模拟法：模拟现实中最有可能发生的异常作答反应，根据测验的理论框架如项目反应理论(item response theory, IRT)模型通过项目参数模拟一批异常被试进行标记，再与正常被试混合训练模型，这样的方法虽然只能在一定程度上代表现实中的情况，但是我们可以通过尽量的泛化数据(尽可能增加所模拟的异常被试得分的随机性)等手段来贴近现实。如 Zopluoglu (2019)在项目预知的研究展望中提出使用多层 IRT 模型拟合到所调查的数据集，并模拟具有项目反应和反

应时间的已标记数据(基于调查数据集中的相同项目参数)嵌入数据集中来训练模型; Zhu 等人(2022)也同样提出通过认知诊断理论模拟被试在真实情景中最可能出现的异常情况(作弊、随机作答、睡眠效应)来训练模型。

(2)手动标记法：使用检验力高、不依靠理论假设的非参数统计量来标记异常受试者或者根据受试者的明显异常特征来手动标记异常受试者(比如将极快的正确作答标记为可疑), 再将其嵌入训练集进行训练, 这样模型就可以根据明显异常的被试特征泛化到与其类似特征的被试, 这个方法对于统计量或者异常特征的明显程度要求比较高, 因此在使用时仅建议标记极端受试者。Meng 和 Ma (2023)使用鉴别考生作答相似性的 RSI 指标先手动对部分受试者进行标记, 再通过标记的数据进行训练来检测作弊抄袭考生; Pan等人(2022)提出根据被试的作答速度与是否答对的信息标记一部分极端的异常被试, 使用自训练算法训练分类器从而识别更多的未标记数据。

(3) SMOTE 方法：对少数类样本进行插值, 生成一些与已有的少数类样本相似但略有变化的合成样本, 以增加少数类样本的数量, 可以提高模型对少数类别的预测能力, 从而提高整体模型性能, 主要用来解决数据不平衡的问题。在项目预知的研究中最常被用到的计算机认证考试数据集中(Cizek & Wollack, 2017), 被标记为异常的受试者仅占 2.81%, 在这种情况下对多数类进行低采样或放弃非作弊者样本都是不合适的, 失衡比例的样本对模型的训练将产生消极的影响。例如, Zhou 和 Jiao (2023)通过使用 SMOTE 增加来合成作弊受试者数据。这意味着当我们仅拥有小部分已标记的异常样本时, 我们可以通过合成少数群体过度采样技术适当增加异常样本, 该方法在数据本身噪点较大的情况下要谨慎使用。

这些方法都是通过在原始数据集上添加更多异常样本来增强数据, 但是都有一定的使用限制, 因此在使用时需谨慎。

5.4　初始数据处理

机器学习数据集的处理一般要经历三个阶段：缺失值插补、数据编码以及特征缩放, 在测验数据中经常会发现缺失值, 例如一张调查问卷中部分被试有相当数量的项目都未作答。在机器学习领域, 通常假设所有变量都包含合理的值, 但如果存在缺失值则会不同程度的影响模型效果。一种比较直接的方法是舍弃缺失值较多的某个样本或某个特征, 这仅适用于数据量较大的情况, 在处理时应该尽量通过一些插补的方法减小数据损失, 例如使用中位数、平均数, 或者随机森林插补法(Stekhoven & Bühlmann, 2012), 也可以酌情将其全部输入为 0 来保留特征维度(Zhen & Zhu, 2024)。在数据编码方面, 数值型数据对大部分算法都比较友好, 例如我们经常用到的项目得分或者反应时等都是数值型数据, 多数时候需要将非数值型数据进行编码, 在测验安全领域中我们最常用到序数编码和独热编码, 前者适用于可以排序的顺序数据, 例如本科生、硕士生、博士生, 而后者尤其适用于由非顺序类别(如作答选项中的 ABCD、地名等)数据(Zopluoglu, 2019)。在特征缩放方面, 分数、被试作答次数等数据一般不会差距太大, 但若某些变量普遍出现很大差距, 应使用标准化方法将其缩放到可比较的尺度。

5.5　输入特征的选择

我们可以通过增加特征数量来增强特征空间, 我们所选择特征的质量对机器学习模型的性能有重大影响。在原本特征的基础上添加新特征可以增强输入数据的特征表示, 从而提高模型性能(Heaton, 2016)。Zopluoglu (2019)通过将项目反应进行独热编码来扩展特征空间。Jiao 等人(2023)在堆叠学习算法中, 通过在特征空间中添加基于项目反应和反应时间的个人拟合统计量作为特征变量训练模型, 被证实能有效提高作弊检测的准确率。在训练模型过程中, 如果出现由于输入特征较少或者特征质量较差导致模型效果差的情况, 我们可以选择增加特征变量, 提高模型效果, 比如在仅有项目反应作为特征时, 我们可以有针对性的加入一些与目标相关性强的统计量或人口统计学信息等新特征。

虽然我们建议在调试模型时多尝试一些特征变量, 但是模型性能并不是靠特征数量而是靠适量取胜, 在训练模型时我们需要不断筛选合适的特征, 这个过程称为特征选择, 其目的是为了减少训练时间、提高模型精度并防止过拟合(Chen et al, 2020), Zhou 和 Jiao (2022)在研究中使用了过滤法、包装法和嵌入法分别进行特征选择。在测验安全领域, 受试者反应时普遍被认为是能有效筛选异常的特征, 它可以捕捉到特定事件的速度

或持续时间, 在进行特征选择时可重点关注, 除了直接将其作为输入特征以外, 还可以计算反应时与其他特征的关系, 如反应时残差, 反应时与分数的相关性。

6 问题与展望

如何侦查各种测验中的异常被试已经成为许多教育机构和考试公司的重要问题, 异常被试不仅威胁着测验分数的可靠性与解释性, 也给测验的名誉带来严重损害, 不利于心理和教育测量的发展。目前, 研究者们提出了许多有效的方法对其进行识别, 与此同时, 各种过程数据的出现给传统的以统计量为主的评估框架带来了机遇和挑战, 机器学习在测验安全领域的研究也不断扩而充之, 但研究基本都集中在国外, 其中使用到的算法非常丰富, 应用的场景也十分广泛。这些研究为我们今后在测验安全领域中使用机器学习方法识别异常被试提供了参考, 可以将机器学习与传统的统计量研究高度结合, 更好的检测异常受试者。此外, 当前机器学习的文本挖掘技术已经充分运用到检测学术上的剽窃抄袭(Foltýnek et al., 2019)以及检测学术论文或大型文字任务中的人工智能生成内容(Taloni et al., 2024), 由此看来, 使用文本挖掘技术检测异常受试者是非常有潜力的。总的来说, 机器学习作为一种在心理与教育测量领域新兴的方法, 在当前研究中充分体现了其优势, 但也有些局限性需注意, 如受限于已标记数据的数量和质量、模型的可解释性有待提高、实验的可重复性有待加强等。需要特别注意的是, 尽管在研究中取得了一定成果, 但是这仅能作为一种统计学意义上的辅助标记手段, 在现实中要谨慎对待。当前机器学习在国内的测验安全领域研究还比较稀少, 对于其方法改进的理论研究或者对其实际应用的实践研究都有待探索。现针对机器学习在当前研究中存在的问题以及未来可能的研究方向提供一些建议, 以供研究者参考。

6.1 基于机器学习的个人拟合研究

当前测验安全领域大多数研究者都在关注如何找出作弊考生, 在当今高竞争力的环境下, 作弊确实是最威胁筛选性考试(例如高考、各种职业资格证考试等)也是最值得关注的异常类型, 但是其他的异常类型如随机作答、疲劳效应、创造性作答等却严重影响着学生的学业评估和测验的个人拟合。这方面的研究主要集中在个人拟合统计量上, 机器学习的研究却比较稀少, 但是考虑到学业测试中一套试题经常会重复使用两三年甚至更久, 以及项目得分和反应时随着题目难度变化的规律, 使用机器学习进行个人拟合研究是十分有潜力的, 因此 Zhu 等人(2022)提出了基于神经网络的个人拟合检验方法, 该方法针对课堂的短测验取得了良好的效果。根据该研究的思路, 可以先获取一批能力分布较为均匀的无污染数据, 再根据测验的理论框架估计项目参数, 生成正常被试与不同类型的异常被试作为训练数据集对模型进行训练, 在训练过程中可以使用各种监督学习模型或是集成学习模型, 待模型训练好后便可重复使用, 从而省去面对新数据重新进行参数估计和模型训练的步骤。这种方法在面对某些统计量方法检测效果较差或者没有对应测验类型的统计量可用时能发挥很大作用, 在各种理论框架和测验类型下都有较大的研究空间。

6.2 基于多模态数据的机器学习测验安全研究

目前测验安全领域的机器学习研究仍处于起步阶段, 多数研究都基于常规测验数据(项目反应、项目得分、反应时等), 然而除了这些数据仍有许多其他类型的数据值得我们去关注和研究: (1)统计量: Zhou 和 Jiao (2022, 2023)、Jiao 等人(2023)的研究结论充分表明了将与检测目的高度相关的统计量纳入输入特征后模型效果得到了明显提升, 例如在检测抄袭的研究当中纳入高检验力的抄袭统计量作为输入特征, 当前有许多研究的输入特征中都包含了统计量, 这也一定程度上代表着模型性能反映了所选统计量的检验力, 而不是机器学习技术, 进一步说明了统计量与机器学习方法的互补性; (2)图片与视频: Hussein 等人(2022) 使用五种不同类型的已知作弊特征在作弊视频的帧层面对动作检测任务进行了研究, 从而建立检测作弊的模型; (3)计算机活动日志: Tang 等人(2023)、Alsabhan (2023)等研究者利用被试在计算机上的行为数据(点击流)进行时间序列建模来检测被试是否出现了异常行为; (4)生物特征信息: Ullah 等人(2019)提出了一种电子考试监控系统, 使用眼动追踪器和指纹读取器对被试在屏幕上的总时间和他们离开屏幕的频率进行检测, Rodríguez-Villalobos 等人(2023)开发了一种评估头部位置和时间延迟的系统, 讨论了作弊行为

与被试头部相对于计算机屏幕的位置变化之间的高度统计相关性。可以发现, 当前研究不断地向更丰富的数据类型探索, 其中的大量信息尤其是生物特征信息是无法造假的, 如果成功将其与常规测验数据结合进行研究, 将大大提高研究结果的准确度和可信度, 这对于重要考试来说是非常有必要的。我们可以尝试将这些多模态数据融合起来对监督学习模型进行训练, 或者是进行其他机器学习研究, 在模型开发过程中会涉及到不同类型数据如何转换为有效输入特征的问题, 同时对模型的选择和使用也需要更多考量, 例如, 如何将常规数据和图片、眼动数据等同时作为输入特征? 使用哪些模型可以更好地利用不同的数据? 这将是一个非常有挑战性和前景的话题。

6.3　基于生成对抗网络的测验安全研究

本文在 5.3 节提到了三种获得异常被试样本的方法, 实则在机器学习领域仍有许多方法可以获得更加贴近真实的数据, 例如深度学习中的生成对抗网络(generative adversarial network, GAN), GAN 由 Goodfellow 等人(2020)提出, 该网络由生成模型和判别模型构成, 生成模型不断捕捉真实数据的分布, 判别模型判断输入数据是真实数据还是生成器所生成的数据, 二者相互博弈训练, 最终使生成模型学习到最逼近真实数据的分布。GAN 既可以生成真实数据, 也可以用来进行异常检测。Zenati 等人(2018)首次使用 GAN 来识别网络数据集中的异常入侵数据; Di Mattia 等人(2019)对异常检测的 GAN 模型进行了比较和分析。Zopluoglu (2019)提到在测验安全的研究中, GAN 可以在数据集量少不足的情况下, 根据这部分少量的数据集的特征来生成更多新数据集从而扩充训练数据, 想实现这个功能需要: (1) 收集正常数据, 训练一个包括生成器和判别器的标准 GAN, 以生成与正常数据分布一致的数据; (2)生成异常数据, 对生成的正常数据进行小幅度的扰动, 添加噪声或改变某些特征值, 从而生成异常样本, 可以参考 5.3 节; (3)使用正常数据和异常数据训练反向 GAN, 此时生成器的目标是生成异常数据, 判别器的目标是区分正常数据和生成的异常数据; (4)评估生成的异常数据的质量, 调整参数, 提高生成数据的质量。如果想要进一步实现异常检测, 则需将待检测数据输入生成器, 生成器会根据训练过的数据重建待检测数据, 再比较生成数据与待检测数据之间的重建误差即可识别异常数据, 这部分类似于前文提到的自编码器。总之, 使用 GAN 进行异常检测的研究仍处于起步阶段, 有较大的研究价值。

6.4　增强研究结果的可解释性

机器学习算法被认为是一种“黑箱方法”, 因为它们更多是由数据驱动的, 涉及到所有输入变量之间的复杂关系, 对建模的过程和结果的解释一直存在争议。研究者认为在测验安全的研究中, 对于被试分类结果的解释非常重要, Zopluoglu (2019)根据极端梯度提升模型的变量重要性图表解析了一部分被试的分类结果, 从这样的解释中我们可以获悉哪些特征对被试被分类为异常的影响最大, 例如某些题目的作答是否正确、某些题目的反应时、某些人口统计学变量, 这对于我们分析某个被试或者了解更多关于被试的背景特征是十分有必要的。同时这也使得我们的研究结论更加丰富和完整。其次, 如果研究侧重于对结果的解释, 可以使用更容易解释的简单模型。例如, 线性回归、决策树等模型相对于深度神经网络等复杂模型具有更好的可解释性。可以通过画特征重要性图以及使用一些专门的解释性技术, 例如局部可解释性方法、全局可解释性方法(Du et al., 2019)。

参考文献

韩丹, 郭庆科, 王昭, 陈雪霞. (2008). 考试抄袭识别的心理测量学研究回顾. *心理科学进展, 16*(1), 175−183.

胡佳琪, 黄美薇, 骆方. (2020). 考试作弊甄别技术的研究进展：个体作弊的甄别. *中国考试*, (11), 32−36.

黄美薇, 潘逸沁, 骆方. (2020). 结合选择题与主观题信息的两阶段作弊甄别方法. *心理科学, 43*(1), 75−80.

刘冬予, 骆方, 屠焯然, 饶思敬, 沈阳. (2024). 人工智能技术赋能心理学发展的现状与挑战. *北京师范大学学报(自然科学版), 60*(1), 30−37.

刘玥, 刘红云. (2021). 心理与教育测验中异常作答处理的新技术: 混合模型方法. *心理科学进展, 29*(9), 1696−1710.

骆方, 王欣夷, 徐永泽, 封慰. (2020). 考试作弊甄别技术的研究进展：团体作弊的甄别. *中国考试*, (11), 37−41.

童昊, 喻晓锋, 秦春影, 彭亚风, 钟小缘. (2022). 多级计分测验中基于残差统计量的被试拟合研究. *心理学报, 54*(9), 1122−1136.

王昭, 郭庆科, 岳艳. (2007). 心理测验中个人拟合研究的回顾与展望. *心理科学进展, 15*(3), 559−566.

徐静, 骆方, 马彦珍, 胡路明, 田雪涛. (2024). 开放式情境判断测验的自动化评分. *心理学报, 56*(6), 831−844.

张龙飞, 王晓雯, 蔡艳, 涂冬波. (2020). 心理与教育测验中异常反应侦查新技术：变点分析法. *心理科学进展, 28*(9), 1462−1477.

钟晓钰, 李铭尧, 李凌艳. (2021). 问卷调查中被试不认真作答的控制与识别. *心理科学进展, 29*(2), 225−237.

钟小缘, 喻晓锋, 苗莹, 秦春影, 彭亚风, 童昊. (2022). 基于作答时间数据的改变点分析在检测加速作答中的探索——已知和未知项目参数. *心理学报, 54*(10), 1277−1292.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Alsabhan, W. (2023). Student cheating detection in higher education by implementing machine learning and LSTM techniques. *Sensors, 23*(8), 4149.

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489−2505.

Arthur, W., Jr., Hagen, E., & George, F., Jr. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior, 8*, 105−137.

Cavalcanti, E. R., Pires, C. E., Cavalcanti, E. P., & Pires, V. F. (2012). Detection and evaluation of cheating on college exams using supervised classification. *Informatics in Education, 11*(2), 169−190.

Chan, K., & Stolfo, J. (1997). On the accuracy of meta-learning for scalable data mining. *Journal of Intelligent Information Systems, 8*(1), 5−28

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data, 7*(1), 52.

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4−19.

Di Mattia, F., Galeone, P., De Simoni, M., & Ghelfi, E. (2019). A survey on gans for anomaly detection. *arxiv preprint arxiv: 1906.11632*. https://doi.org/10.48550/arXiv.1906.11632

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*, 241−258.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM, 63*(1), 68−77.

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys (CSUR), 52*(6), 1−42.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139−144.

Gorgun, G., & Bulut, O. (2022). Identifying aberrant responses in intelligent tutoring systems: An application of anomaly detection methods. *Psychological Test and Assessment Modeling, 64*(4), 359−384.

Heaton, J. (2016). An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016* (pp. 1−6). IEEE.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85−126.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828−845.

Hussein, F., Al-Ahmad, A., El-Salhi, S., Alshdaifat, E. A., & Al-Hami, M. T. (2022). Advances in contextual action recognition: Automatic cheating detection using machine learning techniques. *Data, 7*(9), 122.

Jiao, H., Yadav, C., & Li, G. (2023). Integrating psychometric analysis and machine learning to augment data for cheating detection in large-scale assessment. *OSF*. https://doi.org/10.31234/osf.io/fjz2c

Kamalov, F., Sulieman, H., & Santandreu Calonge, D. (2021). Machine learning based approach to exam cheating detection. *Plos One*, *16*(8), e0254340. https://doi.org/10.1371/journal.pone.0254340

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277−298.

Kim, D., Woo, A., & Dickison, P. (2016). Identifying and investigating aberrant responses using psychometrics-based and machine learning-based approaches. In G. J. Cizek & J. A.Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 70−97). New York, NY: Routledge.

Liao, M., Patton, J., Yan, R., & Jiao, H. (2021). Mining process data to detect aberrant test takers. *Measurement: Interdisciplinary Research and Perspectives, 19*(2), 93−105.

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement, 56*(2), 251−279.

Meng, H., & Ma, Y. (2023). Machine learning–based profiling in test cheating detection. *Educational Measurement: Issues and Practice, 42*(1), 59−75.

Pan, Y., Sinharay, S., Livne, O., & Wollack, J. A. (2022). A machine learning approach for detecting item compromise

and preknowledge in computerized adaptive testing. *Psychological Test and Assessment Modeling, 64*(4), 385−424.

Pan, Y., & Wollack, J. A. (2021). An unsupervised-learning-based approach to compromised items detection. *Journal of Educational Measurement, 58*(3), 413−433.

Pan, Y., & Wollack, J. A. (2023). A machine learning approach for the simultaneous detection of preknowledge in examinees and items when both are unknown. *Educational Measurement: Issues and Practice, 42*(1), 76−98.

Ranger, J., Schmidt, N., & Wolgast, A. (2020). The detection of cheating on E-exams in higher education—The performance of several old and some new indicators. *Frontiers in Psychology, 11*, 568825. https://doi.org/10.3389/fpsyg.2020.568825

Ranger, J., Schmidt, N., & Wolgast, A. (2023). Detecting cheating in large-scale assessment: The transfer of detectors to new tests. *Educational and Psychological Measurement, 83*(5), 1033−1058.

Rodríguez-Villalobos, M., Fernandez-Garza, J., & Heredia-Escorza, Y. (2023). Monitoring methods and student performance in distance education exams. *The International Journal of Information and Learning Technology, 40*(2), 164−176.

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement, 82*(1), 29−56.

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46−68.

Stekhoven, D., & Bühlmann, P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112−118.

Taloni, A., Scorcia, V., & Giannaccare, G. (2024). Modern threats in academia: Evaluating plagiarism and artificial intelligence detection scores of ChatGPT. *Eye, 38*(2), 397−400.

Tang, S., Samuel, S., & Li, Z. (2023). Detecting atypical test-taking behavior with behavior prediction using LSTM. *Psychological Test and Assessment Modeling, 65*(2), 76−124.

Thomas, S. L. (2016). *So happy together? Combining Rasch and item response theory model estimates with support vector machines to detect test fraud*. (Unpublished doctorial dissertation). University of Virginia.

Tiong, L. C. O., & Lee, H. J. (2021). E-cheating prevention measures: Detection of cheating at online examinations using deep learning approach--a case study. *arXiv preprint arXiv:2101.09841*. https://doi.org/10.48550/arXiv.2101.09841

Ullah, A., Xiao, H., & Barker, T. (2019). A dynamic profile questions approach to mitigate impersonation in online examinations. *Journal of Grid Computing, 17*, 209−223.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365−384.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*(2), 199−217.

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology, 74*, 577−596.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*, 1−40.

Welz, M., & Alfons, A. (2023). I don't care anymore: Identifying the onset of careless responding. *arXiv preprint arXiv: 2303.07167*.https://doi.org/10.48550/arXiv. 2303.07167

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., & Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection. *arxiv preprint arxiv:1802.06222*. https://doi.org/10.48550/arXiv.1802.06222

Zhen, Y., & Zhu, X. (2024). An ensemble learning approach based on TabNet and machine learning models for cheating detection in educational tests. *Educational and Psychological Measurement, 84*(4), 780−809.

Zhou, T., & Jiao, H. (2022). Data augmentation in machine learning for cheating detection in large-scale assessment: An illustration with the blending ensemble learning algorithm. *Psychological Test and Assessment Modeling, 64*(4), 425−444.

Zhou, T., & Jiao, H. (2023). Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement, 83*(4), 831−854.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*(1), 1−130.

Zhu, Z., Arthur, D., & Chang, H. H. (2022). A new person-fit method based on machine learning in CDM in education. *British Journal of Mathematical and Statistical Psychology, 75*(3), 616−637.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 5*(5), 363−387.

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). *Educational and Psychological Measurement, 79*(5), 931−961.

Application of machine learning methods in test security

GAO Xuliang, LI Ning

(*School of Psychology, Guizhou Normal University, Guiyang* 550025, *China*)

Abstract: The post hoc detection of test security has traditionally relied on statistics, but emerging machine learning methods offer enhanced detection performance. To advance the field of test security, we proposed a review of the research literature, categorizing the methods into three major categories: supervised learning, unsupervised learning, and semi-supervised learning. Each of these major categories was further subdivided into three subcategories: ensemble learning, deep learning, and transfer learning. The study elucidated the distinctive attributes of diverse machine learning methodologies, provided practical recommendations for data acquisition and processing, and outlined strategies for input feature selection. Finally, prospective avenues for future research were identified, including machine learning-based person-fit research, machine learning test security research utilizing multimodal data, test security research employing generative adversarial networks, and the interpretability of research results.

Keywords: machine learning, psychological tests, educational tests, test security, statistics