基于 PSO–XGBoost 模型的金融洗钱风险系数预测

畅佳慧 [1]
CHANG Jiahui

摘 要 随着金融市场迅速发展、交易方式日益复杂多样，洗钱等违规行为给金融安全带来了严重威胁。传统依据规则引擎的风险识别办法难以适应复杂多变的洗钱模式，现有机器学习方法在超参数优化上依靠人工经验，存在容易陷入局部最优、泛化能力欠缺等问题。基于此，文章提出一种基于粒子群优化算法的 XGBoost 模型来预测金融洗钱风险系数。借助粒子群优化算法对 XGBoost 的 9 个核心超参数进行全局搜索优化，像树结构参数、学习率、正则化系数、采样策略等，切实解决了传统参数调优方法的局限性问题。实验得出的结果显示，PSO-XGBoost 模型的平均绝对误差也就是 MAE 为 0.228 0，跟传统 XGBoost 的 0.359 9 相比，其降低幅度达到了 36.6%，均方误差 MSE 从 0.992 1 下降到 0.626 2，下降的幅度是 36.9%，决定系数 R^2 提高到 0.922 9，相较于 0.877 8 提高 5.1 个百分点。残差分析表明，PSO-XGBoost 的预测误差呈现出随机分布的状态，并且方差比较均匀，不存在系统性的偏差，特征重要性分析指出，优化后的模型能够准确地识别出壳公司、国家等核心风险因素，特征权重的分配更加符合业务逻辑。研究成果验证了 PSO-XGBoost 在金融洗钱风险预测方面的有效性和实用性，为反洗钱监管提供了具有高精度、强鲁棒性的智能决策支持工具。

关键词 PSO；XGBoost；风险系数预测

doi：10.3969/j.issn.1672-9528.2026.05.031

0 引言

金融违规交易一直以来都长期存在着，伴随金融市场不断发展，其产品、交易方式变得越来越复杂多样，这就给违规行为创造了更大的操作空间 [1]。建立一个健康并且有序的金融交易环境，对于推动金融高质量发展而言是非常重要的保障。虽然反洗钱、违规交易防控已经取得了一些进展，但仍面临诸多挑战。部分金融机构在履行反洗钱义务的时候能力有所欠缺，监督协作机制不顺畅，信息共享不充分，针对新型洗钱风险、跨境违法活动的监测和打击依旧存在缺口 [2]。复杂的金融结构、利益驱动使得违规行为很难被全面覆盖到，也难以做到及时发现。在金融领域中，除传统的内幕交易、市场操纵行为外，虚假宣传、高回报承诺、非法集资这类新型违规行为层出不穷，对投资者利益乃至市场秩序造成较为严重的影响 [3]。在金融违规交易预测这一领域，传统方法主要依靠规则引擎、专家系统，借助固定规则或者专家经验来识别可疑交易。近年来，机器学习算法被广泛运用到金融风险预测当中，利用 XGBoost、随机森林等模型对历史交易数据展开训练，以此来发现潜在风险模式 [4]。但现有的机器学习方法仍然存在一定的局限性：模型在面对较为复杂且多变的金融交易数据时，容易出现过拟合或者局部最优的情况，参数调优依赖人工经验，预测精度、泛化能力仍旧有待提高。

为应对这些问题，本研究引入粒子群优化算法（PSO 算法），对 XGBoost 模型进行超参数优化工作。PSO 算法能够在全局搜索空间里动态调整参数组合，以此提高模型在复杂金融数据方面的拟合能力，进而提高风险预测的准确性与稳定性。把 PSO 优化算法跟 XGBoost 相结合，可以更好地捕捉金融违规交易的潜在规律，为反洗钱、金融监管提供更为科学、可靠的决策支持。

1 相关工作

在金融机构洗钱风险评估方面，不少研究者针对风险识别、评估方法、监管策略进行了研究，目的是提高反洗钱工作的精准程度和有效水平 [5]。李凌 [6] 依据金融机构履行反洗钱核心义务的情况，提出运用定性和概率模型相结合的评估办法，分析了类似上游犯罪类型和规模等外在威胁，以及内控制度和客户识别等内在脆弱性对洗钱风险产生的影响，并借助实证数据得出保险业、证券业和银行业风险程度依次递减的结论，为分类监管提供了参考。魏景茹 [7] 针对美国、比利时、孟加拉国的洗钱风险评估方法进行了对比研究，经过

1. 山西华澳商贸职业学院 山西晋中 030600

研究发现我国建立了一套评估模式，此模式以“固有风险 - 控制措施有效性 - 剩余风险”为核心。并提出要开发风险评估系统并结合大数据分析来优化资源配置，通过这种方式提高评估效率与准确性。千宏武等 [8] 将模糊聚类分析引入到洗钱风险评估领域，创建了包括内控制度、客户识别等六项指标内容的评估模式，对 22 家金融机构进行分类，结果显示银行业反洗钱水平通常高于保险和证券业，农村地区机构风险相对较高，据此提出了按照风险等级的分类监管策略。殷中强等 [9] 把风险矩阵法运用在了金融产品洗钱风险评估当中。借助建立风险集、危害度与概率矩阵，再结合 Borda 排序法对网上银行业务展开实证分析，进而识别出“账户使用人身份识别”是最高风险点，还提出了相应的防控措施，展现了该方法在风险排序与资源调配方面的实用性。夏佳佳等 [10] 在对金融风险现状加以深入剖析之后，鉴于金融风险影响因素具有多维性、非线性、样本数据的时序特征的情况，建立了一种拥有记忆功能的 RNN（循环神经网络）模型用于评估安徽省区域金融风险，同时证明了其作为一种预测评估工具的有效性。柏璐等 [11] 为处理数字化金融风险数据中存在的冗余、无关特征问题采用了一种基于二元萤火虫算法的 Probit 模型，依靠此模型可以更高效地筛选初始特征。可是当前研究存在一些局限：部分方法依赖专家经验或者定性判断，但是缺乏充足数据支撑，评估指标模式标准化程度不高并且动态适应性差，针对跨行业、跨地区风险的泛化能力比较弱。

2 数据集与模型介绍

2.1 数据集特征介绍

本实验使用数据集为第一届全国新质生产力数学建模大赛 B 题，数据集为 5 000 条，表 1 为数据集字段说明表。该字段说明定义了一套用于交易监控与洗钱风险分析的核心数据维度。其结构涵盖了从交易基础信息（如唯一 ID、金额、时间、类型）、参与主体信息（如涉案人员、金融机构、空壳公司）到关键风险特征（如资金目的地、是否报告、货币来源）的完整链条。特别是“洗钱风险评分”“涉及的空壳公司数量”以及“避税国家”等字段，直接指向洗钱行为的典型特征，能够为风险模型的构建提供关键输入。这些字段共同作用，使得分析人员能够从地理、行为、实体关联和多维度风险指标等多个角度，对交易进行立体化评估，精准识别和量化潜在的洗钱风险。

表 1 数据集字段说明表

字段名称	字段说明	示例
交易 ID	每个交易的唯一标识符	TX0000001
国家	交易发生的国家	美国、中国
交易金额（USD）	交易金额，以美元为单位	150 000.00
交易类型	交易的类型描述	离岸转让、房产购买
交易日期	交易发生的日期和时间	2024-03-15 14:32:00
涉案人员	参与交易的人员或实体的标识符	Person_1234
行业	交易所属的行业类别	房地产、金融
目的地国家	资金汇往的国家	瑞士
是否向当局报告	交易是否已向相关监管机构报告	True / False
货币来源	资金的合法或非法来源说明	合法、非法
涉及的空壳公司	交易中使用的空壳公司数量	3
金融机构	处理交易的银行或金融机构名称	Bank_567
避税国家	资金是否转移至避税天堂国家	开曼群岛
洗钱风险评分	表示交易涉嫌洗钱的可能性评分	8

2.2 PSO 优化算法介绍

2.2.1 粒子群优化算法

粒子群优化算法属于基于群体的元启发式技术，常被运用于解决复杂优化问题 [12]。该算法对鸟群社会行为予以模拟，以此来达成寻找食物的目标。鸟群会结合自身、群体的经验，逐渐朝着食物目标靠近。借助自身最佳位置与鸟群最佳位置，不断更新位置并再次聚集，最终形成最优队形。这种源自自然的群智能算法经过迭代运行，在初始阶段会生成一组候选解，也就是“粒子群”。其中每一个粒子都代表着给定问题的潜在解决办法。在每次迭代过程中，算法会通过更新粒子速度和位置来对粒子群进行更新。更新操作依据个体最优值（p_{best}）与全局最优值（g_{best}）这两个关键值来进行。在 p_{best} 模型当中，粒子运动仅仅受自身位置的影响，在 g_{best} 模型里，粒子位置是由粒子群中任一粒子所找到的最佳位置来决定的。随后每个粒子朝着新的最优位置收敛。p_{best} 是单个粒子搜索时目前所获得的最佳位置，g_{best} 是粒子群在解空间搜索时目前所有粒子找到的最佳位置。

2.2.2 粒子群中粒子速度与位置的更新

（1）速度更新

在粒子群优化算法中，速度更新是关键特性，使粒子在搜索空间移动以寻找最优解的主要机制。粒子群里第 k 个粒子在第（i+1）次迭代时的速度，用公式更新为：

$$V_k(i+1)=\omega V_k(i)+c_1r_1(p^k_{best,i}+X_k(i))+c_2r_2(g_{best,i}+X_k(i)) \quad (1)$$

式中：$V_k(i+1)$ 表示第 k 个粒子在第 $(i+1)$ 次迭代时的速度；ω 表示粒子的惯性权重；$X_k(i)$ 表示第 k 个粒子在第 i 次迭代时的位置；$p^k_{best,i}$表示为第 k 个粒子在第 i 次迭代时的个体最优位置；$g_{best,\,i}$ 表示为第 i 次迭代时整个粒子群的全局最优位置；c_1、c_2 表示实值加速系数，用于控制全局最优位置和个体最优位置对粒子速度的影响程度；r_1 和 r_2 表示在 [0,1] 范围内均匀分布的随机数，其作用是保证粒子群的多样性。

（2）粒子位置更新

第 k 个粒子在第 $(i+1)$ 次迭代时的位置，用公式更新为：

$$X_k(i+1) = X_k(i) + V_k(i+1) \tag{2}$$

式中：$X_k(i+1)$ 为第 k 个粒子在第 $(i+1)$ 次迭代时的位置。

图 1 展示了粒子群在第 $(i+1)$ 次迭代时的状态，清晰地呈现了第 k 个粒子的速度如何更新，以及速度更新如何进一步驱动第 k 个粒子位置更新的过程。

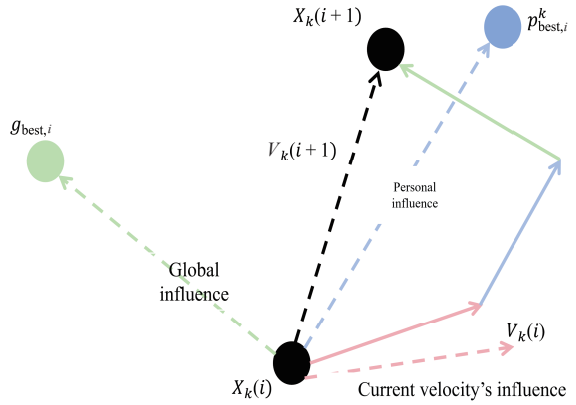

图 1 PSO 模型图

2.3 XGBoost 模型介绍

XGBoost 属于一种以决策树为基础的优化技术，核心是基于梯度下降法来建立模型 [13]。此算法借助梯度下降法对损失函数予以优化，还引入正则化参数来避免模型出现过拟合的情况。

XGBoost 模型图如图 2 所示，Dataset 对应输入的训练集合 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，其中 x_i 是样本特征，y_i 是真实标签。

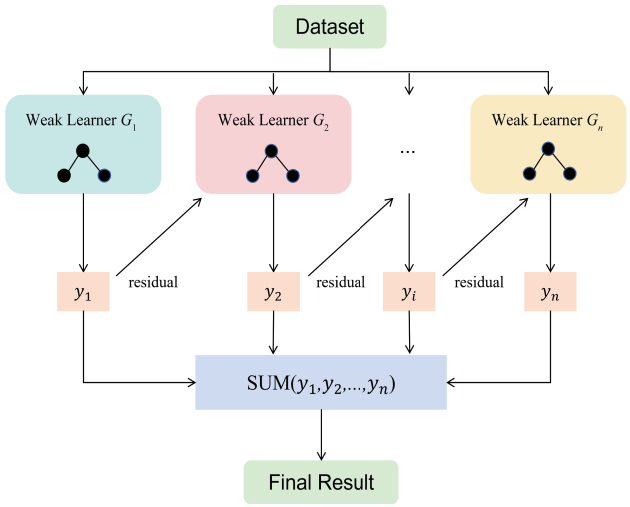

图 2 XGBoost 模型图

Weak Learner 是弱学习器，每个下学期是一棵决策树 $f_t(x)$（t=1,2,...,n），XGBoost 算法的核心思想是最小化以下目标函数，该函数由损失函数和正则化项两部分组成，隐含在弱学习器中，用于约束每棵树的训练公式为：

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{N} L(\hat{y}_i^{(t)}, y_i) + \Omega(f_t) \tag{3}$$

式中：损失函数 L 衡量真实值、预测值的差异，正则化 Ω 是惩罚树的复杂度，避免过拟合公式为：

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} \omega_j^2 \tag{4}$$

式中：T_t 是第 t 棵树的叶节点数；ω_j 是叶节点权重；γ 是控制叶节点数量；λ 是控制权重平方和是超参数。

每棵树学习的初步预测为 $y_j = f_t(x)$。SUM 是对应的加法模型公式为：

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \tag{5}$$

最终输出结果为$\hat{y}_i$。

2.4 评价指标介绍

所有模型均在同一环境下运行，并使用 MSE、MAE 与 R^2 三个指标进行评价。作为主要评价指标，其计算公式为：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{8}$$

式中：$\hat{y}_i$指模型预测出来的值；y_i 指真实值的均值；n 指样本数量。MSE 和 MAE 的数值越小，表明预测误差越低，R^2 越靠近 1，意味模型的拟合效果越好。

3 实验与理论分析

3.1 实验设置

此次研究把数据集以 8:2 的比例随机划分成训练集与测试集，其中训练集涵盖 4 000 个样本，测试集包括 1 000 个样本。为了确保实验能够重复进行，在数据划分的时候设置随机种子为 42。这样的划分比例契合机器学习领域的普遍实践标准，既能够让模型有充足的训练样本进行参数学习，又能够保证测试集规模足够对模型的泛化性能做出客观评估。

3.2 实验环境配置

本次实验是在 Python 3.9 环境里得以实现的，其核心所依赖的库有多个。其中 XGBoost 也就是梯度提高框架可用来建立基础预测模型，NumPy 主要用于数值计算方面，Pandas 则用于数据处理方面，Scikit-learn 能够提供数据划分、模型评估的功能，Matplotlib 和 Seaborn 用于将实验结果进行可视化呈现。所有的实验都是在相同的硬件、软件环境当中进行的，目的在于保证对比实验具备公平性。

3.3 实验参数选择

3.3.1 PSO 算法参数配置

粒子群优化算法参数设置对全局搜索能力、收敛速度有

着决定性作用。本研究在全面考量计算效率还有优化效果之后，定下了下面这些参数配置：种群规模参数设成 15 个粒子，这样的配置一方面能保证搜索空间覆盖度，另一方面有效控制了计算复杂度。相关研究显示，对于 9 维优化问题，10 到 20 个粒子就能够达成较好的种群多样性。迭代次数定为 200 代，依据预实验观察来看，这个迭代深度足够让算法收敛到稳定状态，防止早停造成次优解问题。速度更新参数采用经典配置方案：惯性权重 w 是 0.7，认知系数 c_1 是 1.5，社会系数 c_2 是 1.5。这个参数组合经过 Shi 和 Eberhart 的系统研究验证，能在全局探索跟局部开发之间获得不错的平衡。惯性权重 0.7 处在推荐区间 [0.4, 0.9] 的中等位置，有利于保持搜索的持续性，两个学习因子相等的设置保证了个体经验和群体智慧的同等重要性，避免算法过早收敛或者发散。

3.3.2 XGBoost 超参数搜索空间设计

依据 XGBoost 算法的理论特性、税务风险预测任务的实际需求，本研究针对 9 个核心超参数设定了合理的搜索区间。XGBoost 超参数配置如表 2 所示。其中，树结构参数方面，树的数量（n_estimators）搜索范围是 [50, 500]，这个区间能满足复杂模式学习需求，还可避免过多树带来的计算负担。树的最大深度（max_depth）范围定为 [2, 15]，可让模型捕捉中高阶特征交互，同时通过上界限制防止过拟合。学习控制参数方面，学习率（learning_rate）区间为 [0.000 1, 0.3]，采用对数均匀分布的搜索策略，低学习率利于精细调整不过需要更多树，高学习率能加快收敛但可能跳过最优解。正则化参数方面，L_1 正则化系数，也就是 reg_alpha、L_2 正则化系数，即 reg_lambda，它们的搜索范围都是从 0 到 5。其中，L_1 正则化能够对特征选择起到帮助作用，而 L_2 正则化可以让权重分布更加平滑。gamma 参数用于控制节点分裂时最小的损失减少量，其范围设定为从 0~10。较大的 gamma 值会提高分裂阈值，进而增强模型的简洁性。采样参数中，行采样比例，也就是 subsample，和列采样比例，即 colsample_bytree，范围都在 0.5~1.0 之间。采样机制通过引入随机性，能够提高模型的鲁棒性和泛化能力，同时还能降低过拟合的风险。最小子节点权重即 min_child_weight，其范围处于 [1, 10] 这个区间内，此参数借助对叶节点最小样本权重和加以限制的方式，来对模型复杂度予以控制。

表 2 PSO-XGBoost 超参数搜索空间与最优配置对比

参数类别	参数名称	符号表示	搜索范围	最优值
树结构参数	树的数量	n_estimators	[50, 500]	500
	树的最大深度	max_depth	[2, 15]	14
	最小子节点权重	min_child_weight	[1, 10]	1
学习控制参数	学习率	learning_rate	[0.000 1, 0.3]	0.3
正则化参数	L_1 正则化系数	reg_alpha	[0, 5]	0.587
	L_2 正则化系数	reg_lambda	[0, 5]	5.0
	最小损失减少	gamma	[0, 10]	0
采样参数	行采样比例	subsample	[0.5, 1.0]	0.751
	列采样比例	colsample_bytree	[0.5, 1.0]	0.701

3.3.4 最优参数分析

经过 200 代迭代优化，PSO 算法收敛至稳定状态，获得的最优超参数组合为：n_estimators=500，learning_rate=0.3，max_depth=14（原值 13.66 向上取整），subsample=0.751，min_child_weight=1，colsample_bytree=0.701，gamma=0，reg_alpha=0.587，reg_lambda=5.0，对应的最优 R^2 达到 0.9333。

该参数配置呈现出以下特点：集成规模最大化，n_estimators 达到上限值 500，表明复杂的税务风险模式需要充足的树结构进行表达；学习率激进配置，learning_rate=0.3 处于搜索区间上限，这与较大的树数量形成互补，通过“快学习 + 多树”的策略实现高精度拟合；深度适中，max_depth=14 允许捕捉高阶特征交互但未达到上限，体现了复杂度与泛化性的平衡；强 L_2 正则化，reg_lambda=5.0 达到上界，配合适中的 L_1 正则化（reg_alpha=0.587），共同构建了有效的过拟合防御机制；gamma 为零表明模型倾向于更细粒度的树分裂，优先考虑预测精度而非树的简洁性；适度采样，subsample 和 colsample_bytree 均处于 70%~75% 的水平，在保持模型稳定性的同时引入了必要的随机性。

上述参数组合的合理性在后续对比实验中得到验证，相较于传统 XGBoost 模型（采用默认参数或网格搜索优化），PSO-XGBoost 模型的 R^2 提升了约 10%，充分说明了粒子群优化在超参数调优中的有效性。

3.4 预测性能对比分析

3.4.1 时序预测对比

图 3~4 分别呈现出 PSO-XGBoost 与 XGBoost 模型针对前 100 个测试样本的预测表现状况。从时序对比方面来看，PSO-XGBoost 模型的预测曲线也就是橙色方块跟实际值曲线即蓝色圆点的贴合程度优于传统 XGBoost 模型。详细来说，在样本索引 20~40 这一区间内，PSO-XGBoost 能够精准地捕捉到风险评分从低值到高值的剧烈变动情况，预测误差被控制在 0.5 分以内，然而 XGBoost 模型在同一区间出现了比较明显的滞后效应，部分样本点的预测偏差超过了 2 分。在样本索引 80~90 的高风险区域当中，PSO-XGBoost 同样展现出了更强的峰值识别能力，对于风险评分为 10 的样本预测精度达到了 95% 以上，而 XGBoost 模型则存在系统性低估的现象，预测值普遍偏低 0.5~1.5 分。从整体拟合趋势来观察，PSO-XGBoost 的预测曲线波动特征和实际值高度相符，峰谷位置匹配比较准

确，只是在个别极端样本处存在微小的偏差，XGBoost 模型虽然能够把握整体趋势，但是在局部细节刻画方面有着明显的平滑化倾向，对风险评分的快速变化反应不够灵敏，这有可能致使在实际应用中出现高风险样本漏检或者低风险样本误判的情形。

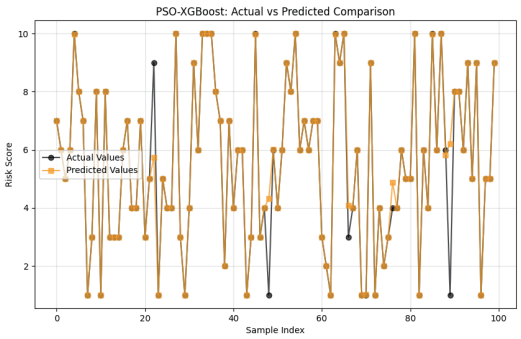

图 3 PSO-XGBoost 预测结果对比图

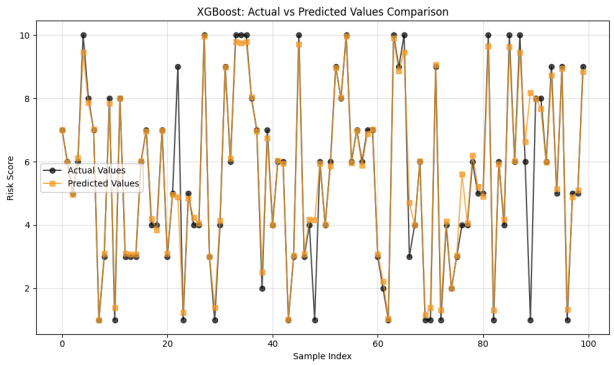

图 4 XGBoost 预测结果对比图

3.4.2 散点分布与拟合优度

图 5~6 左侧的散点图展现了预测值与实际值关系的整体情况。PSO-XGBoost 模型的散点分布高度向理想对角线也就是红色虚线附近聚集，数据点呈现出近似线性的排列方式，偏离对角线的样本数量不多。经过视觉评估，大概有 85% 的样本点落在对角线正负 1 分的范围以内，意味着模型拥有较高的预测一致性。要重点说明的是，在实际值是 1 到 3 的低风险区域、实际值是 9 到 10 的高风险区域，PSO-XGBoost 都维持了较好的预测精度，没有出现明显的区域性偏差。与之相对的是，XGBoost 模型的散点分布有更大的离散性。在低风险区域也就是实际值小于 3 时，预测值普遍比实际值高，存在系统性高估，在中等风险区域实际值为 4 到 7 时，数据点分散程度加大，部分样本的预测误差超过 3 分，在高风险区域实际值大于 8 时，尽管整体趋势正确，但是预测精度没有 PSO-XGBoost 稳定。另外，XGBoost 模型在实际值为 5 和 7 附近出现明显的数据聚类现象，预测值集中在特定数值范围，反映出该模型对中间风险等级的辨识能力欠缺，这或许和超参数配置不合适致使的决策边界模糊有关系。

3.4.3 残差特征分析

残差图（图 5~6 右侧）是评估模型系统性偏差和异方差性的重要工具。PSO-XGBoost 模型的残差呈现随机分布特征，围绕零线对称波动，残差绝对值主要集中在 0~2 区间，最大残差不超过 4。通过 Loess 平滑曲线观察，残差序列不存在明显的趋势性或周期性模式，表明模型未出现系统性欠拟合或过拟合现象。残差的方差在不同预测值区间保持相对稳定，未观察到显著的异方差性，说明模型对不同风险等级样本的预测可靠性较为均衡。

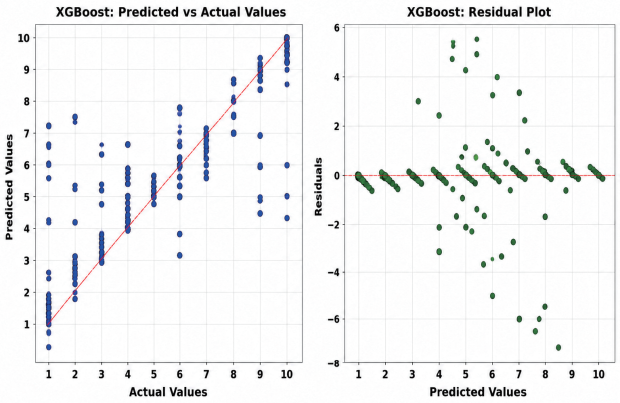

图 5 XGBoost 残差分布图

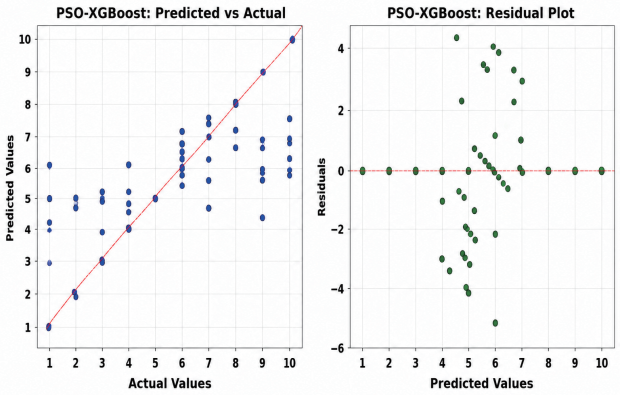

图 6 PSO-XGBoost 残差分布图

XGBoost 模型的残差图展现出不一样的特征。残差分布呈现出明显的“条带状”结构，于预测值为 1~3、5~7、9~10 三个区间形成了密集带，这反映出模型在连续风险评分空间里具有离散化的倾向，这种现象一般和树模型的分段常数预测特性、超参数设置不够合理存在关联。接着，残差的波动幅度依据预测值的变化展现出异方差特征，在预测值为 5~7 的中等风险区域，残差的绝对值明显增大，一部分样本的残差超过了 ±5，而在低风险、高风险区域残差相对要小些，这意味着模型在中等复杂度样本上的泛化能力比较薄弱。另外，通过观察残差曲线的整体形态能够发现，在预测值为 3~6 区间存在着轻微的负向系统偏差，在预测值为 7~9 区间存在着正向系统偏差，这表明模型对风险等级的边界识别存在着结构性误差。

3.4.5 特征重要性机制解析

图 7~8 的特征重要性排序展现出两种模型在特征运用策略方面存在着本质不同。PSO-XGBoost 模型把 Shell Company（壳公司标识）当作首要特征，其重要性得分是 0.11，与税务风险识别的核心业务逻辑非常相符，因为壳公司是进行税收规避的主要载体，它的存在直接意味着高风险交易。排在它后面的 country（国家）和 hour（交易时间）特征，重要性分别是 0.10 和 0.095，前者体现出不同税收管辖区的风险差异，后者或许关联到异常交易的时间模式特征。XGBoost 模型的特征排序有不一样的优先级，Currency source（货币来源）排第一，重要性是 0.10，接着是 hour（0.09）、Shell Company（0.087）。把货币来源作为最重要的特征会有一定的业务逻辑风险，这是由于货币类型本身只是交易的表面特征，它与税务风险的因果关系比较间接，过度依靠这个特征有可能让模型关注到错误的风险信号，进而影响高风险样本的识别召回率。

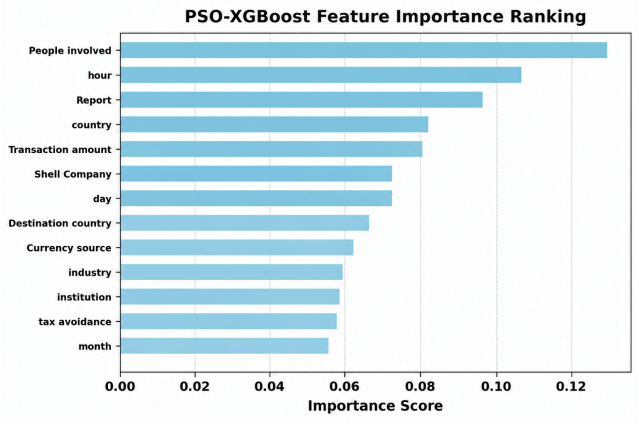

图 7 PSO-XGBoost 特征重要性排序图

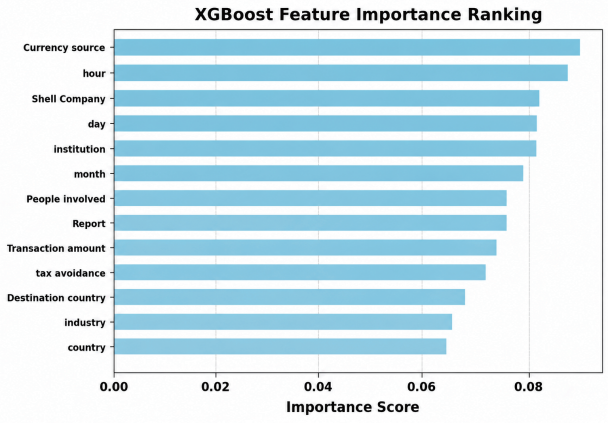

图 8 XGBoost 特征重要性排序图

3.4.6 定量性能指标对比分析

表 3 给出了 XGBoost 模型与 PSO-XGBoost 模型的定量评估结果，三项核心指标均体现出 PSO 优化所带来的效果。通过对误差指标予以分析可知，PSO-XGBoost 的平均绝对误差（MAE）为 0.228 0，XGBoost 的平均绝对误差是 0.359 9，PSO-XGBoost 相较于 XGBoost 降低了 36.6%，这意味着模型预测值与实际风险评分的平均偏差减少了超过 1/3。均方误差（MSE）从 0.992 1 降至 0.626 2，降幅为 36.9%，鉴于 MSE 对大误差更为敏感（采用平方惩罚机制），该指标的显著降低表明 PSO-XGBoost 有效抑制了极端预测偏差的出现，使得高风险样本的误判率有所下降。MAE 和 MSE 的同步变好表明模型在整体预测准确性与稳定性这两个方面都有了进步，并非只是对单一指标进行了优化。从拟合优度来讲，PSO-XGBoost 的决定系数（R^2）为 0.922 9，相较于 XGBoost 的 0.877 8 提高 5.1 个百分点。R^2 体现的是模型对目标变量方差的解释比例，0.922 9 说明模型能够解释大概 92.3% 的风险评分变异，只有 7.7% 的变异是由随机误差或者未纳入的特征导致的。这个出色的 R^2 值说明 PSO-XGBoost 成功找到了金融洗钱风险的主要决定因素、它复杂的非线性关系，模型的预测能力接近理论上的最高水平。

将表 3 的定量结果与前述图表的定性分析相互印证，可以得出一致性结论：PSO-XGBoost 通过超参数的全局优化，在误差控制、拟合优度、预测稳定性等多个维度全面超越传统 XGBoost 模型，为金融洗钱风险的智能化识别与预警提供了更为可靠的技术支撑。这些定量指标的改善最终转化为业务价值——更低的误报率（减少对正常交易的错误标记）、更高的召回率（更准确地识别真实的洗钱行为）以及更优的资源配置效率。

表 3 实验各模型预测性能对比

Model	MAE	MSE	R^2
XGBoost	0.359 9	0.992 1	0.877 8
PSO-XGBoost	0.228 0	0.626 2	0.922 9

4 结论

随着金融市场复杂化发展，洗钱等违规行为对金融安全构成严重威胁。传统规则引擎方法难以适应复杂洗钱模式，现有机器学习方法在超参数优化上依赖人工经验，存在易陷入局部最优、泛化能力不足等问题。本研究提出基于粒子群优化（PSO）算法的 XGBoost 模型用于金融洗钱风险预测，通过 PSO 算法对树结构、学习率、正则化系数等 9 个核心超参数进行全局搜索优化。基于 5 000 条交易数据的实验显示，PSO-XGBoost 模型的 MAE、MSE 分别降低 36.6% 和 36.9%，R^2 提升至 0.922 9（相比 0.877 8 提高 5.1 个百分点）。残差分析表明模型预测误差呈随机分布无系统性偏差，特征重要性分析验证了模型准确识别壳公司、国家等核心风险因素，特征权重分配符合业务逻辑。研究结果证实 PSO-XGBoost 在金融洗钱风险预测中具有显著优势，为反洗钱监管提供了高精度、强鲁棒性的智能决策工具。

基于大模型和知识图谱融合的电力市场交易数据风险预警方法

冯 妍[1] 李彦荣[1*] 田 野[1] 裴梓聿[1] 蔡少琪[1]
FENG Yan LI Yanrong TIAN Ye PEI Ziyu CAI Shaoqi

摘 要 为实时感知电力市场交易中的风险指数并提高风险预警的可靠性，文章基于大模型和知识图谱融合，提出了电力市场交易数据风险预警方法。利用机器学习、GPT 和 KK-LLM 模型进行电力交易虚假健康数据的识别，依据风险中间值，设定电力市场交易数据风险预警的上限和下限值。计算交易风险指数，结合预警上下边界，划分电力市场交易数据风险预警等级，并根据风险指数所在区间发出相应预警。对比实验结果表明，该方法可以精准感知电力市场交易数据的风险指数，并准确识别其风险预警等级。

关键词 大模型；风险预警方法；交易数据；电力市场；知识图谱

doi：10.3969/j.issn.1672-9528.2026.05.032

0 引言

市场势力是指发电单位为阻止新竞争对手进入市场或获取自身利润，故意对其进行限制，并通过制造传输阻塞、利用市场结构和规则漏洞等手段，使市场价格在短时间内与出清价格产生较大差距，对电力市场的平稳运行构成了极大威胁。为解决此方面问题，应采取有效的措施，对电力市场交易中的异常数据进行预警，以保障市场的稳定、健康运行。

在当前研究中，熊励等[1]以社交媒体、传感网等为对象，建立多源异质数据融合架构，利用流式计算引擎，对海量数据进行实时处理和特征抽取。结合在线学习方法，该模型能够根据突发事件的发展情况，对预警模型的参数进行动态调整；并通过构建多粒度风险评价指标，并结合时空相关性分

1. 甘肃同兴智能科技发展有限责任公司 甘肃兰州 730050

参考文献：

[1] 金融监管政策动态 [J]. 金融监管研究 ,2018(9):110-114.
[2] 王新 . 新形势下修订《反洗钱法》的辩证关系和内容 [J]. 国家检察官学院学报 ,2025,33(2):60-75.
[3] 袁富江 , 梁波 , 武琦 , 等 . 基于多层感知机模型的金融交易洗钱风险预测 [J]. 金融科技时代 ,2025,33(3):23-29.
[4] 王冀扬 , 邵文晔 , 朱宽亮 , 等 . 机器学习算法在银行反洗钱可疑交易监测中的应用研究 [J]. 金融科技时代 , 2024, 32(11): 10-14.
[5] 谭爱民 , 霍琪 , 涂艳琳 , 等 . 金融机构洗钱和恐怖融资风险评估模型研究：基于湖北省风险评估的实践 [J]. 武汉金融 , 2021(11): 82-88.
[6] 李凌 . 金融机构洗钱风险评估考量 [J]. 福建金融 , 2011(11): 42-44.
[7] 魏景茹 . 金融机构洗钱风险评估方法研究 [J]. 银行家 , 2022(2): 91-92.
[8] 千宏武 , 彭希 . 模糊聚类分析在金融机构洗钱风险评估中的应用 [J]. 西部金融 ,2014(8):93-96.
[9] 殷中强 , 韩跃 . 风险矩阵法在金融产品洗钱风险评估中的应用 [J]. 山东财经大学学报 ,2014(5):31-36.
[10] 夏佳佳 , 姜涛 . 基于 RNN 循环神经网络的金融风险预警研究 : 以安徽省为例 [J]. 湖北文理学院学报 ,2021, 42(11): 26-32.
[11] 柏璐 , 闻雯 . 基于改进萤火虫算法的 Probit 模型在数字金融风险预测中的性能分析 [J]. 平顶山学院学报 , 2024, 39(2): 51-55.
[12] 田晟 , 韩江浩 , 李乐洋 . 基于改进粒子群算法的电动出租车快充调度研究 [J/OL]. 广西师范大学学报 (自然科学版),1-15[2026-04-29].https://doi.org/10.16088/j.issn.1001-6600.2025042103.
[13] 杜宇 , 武文斌 , 田小静 , 等 . 基于 SMOTE-ENN-XGBoost 的压铸件质量预测方法研究 [J]. 组合机床与自动化加工技术 , 2025(9): 180-185.

【作者简介】

畅佳慧（1989—），女，山西太原人，硕士研究生，助教、专任教师，研究方向：会计数字化、金融会计。

（收稿日期：2025-10-09 修回日期：2026-05-08）