基于LightGBM特征选择算法的财务舞弊识别研究

■ 于　尧
（河北工业大学，天津，300401）

一、引言

选取 2010 至 2019 年我国沪深两市的舞弊及配对企业各 445 家，综合考虑财务及非财务指标，采用 LightGBM 算法进行舞弊识别指标的特征选择，分别运用逻辑回归、支持向量机、随机森林和神经网络等机器学习方法建立舞弊识别模型。结果表明，运用 LightGBM 筛选变量后，舞弊识别效果得到进一步提高，且 LightGBM 算法与逻辑回归方法结合的模型分类效果最佳。

财务舞弊是指为了欺骗财务报告使用者而有意错报或遗漏信息的一种行为，近年来，关于如何有效识别企业的财务舞弊行为这一话题开始受到广泛关注。本文提出采用 LightGBM 算法进行特征选择，并进一步运用逻辑回归、支持向量机、随机森林和人工神经网络 4 种机器学习技术对我国 A 股上市企业财务舞弊进行识别，尝试探索更加有效的财务舞弊识别方法。

二、文献回顾

传统的统计模型存在较强的主观性，且错判率高。近年来，学者们开始使用机器学习技术对企业舞弊行为进行预测。汤凌冰（2016）采用逻辑回归构建了舞弊预警模型。宋晓勇（2016）验证了神经网络分类器在预测财务报告舞弊方面的有效性。曹德芳（2019）发现支持向量机可以很好地识别财务舞弊。Dong（2018）采用决策树方法对财务报表舞弊进行了分类。随着技术发展的日渐成熟，各学者开始使用集成机器学习方法进行舞弊识别，如 Bao（2020）基于原始财务数据利用 Adaboost 方法进行财务欺诈预测。研究发现，基于机器学习技术的舞弊识别结果更加准确。此外，已有研究中的识别特征数量不等，在识别特征较多时，可能导致结果出现过拟合，这时，需要采用特征选择方法消除冗余特征，以提高识别准确率。近些年，夏明等人（2015）运用主成分分析方法进行数据特征降维。王威（2020）使用 Lasso 方法筛选重要指标。陈艺云等人（2018）采用卡方检验方法提取特征指标，以实现企业财务困境的预测。冯炳纯（2019）运用 Relief、Boruta 算法选择识别指标。

综上，目前被广泛使用的机器学习技术主要有逻辑回归、支持向量机等方法，但大多数研究只使用了 1 至 2 种方法对财务舞弊进行识别。此外，特征选择有利于获得更好的识别效果。因此，本文提出使用 LightGBM 集成算法选择初选特征，然后运用逻辑回归、支持向量机、随机森林和神经网络 4 种应用广泛的机器学习技术进行舞弊识别，以寻求更优的识别方法。

三、研究设计

（一）LightGBM 特征选择算法

LightGBM 集成机器学习算法能够通过数据训练得到特征的重要度排序，这是对 GBDT 算法的一种系统改进，性能更加优秀，训练速度更快。GBDT 是一种基于决策树的梯度提升算法，能根据当前决策树损失函数的负梯度提升树的残差近似值，以此拟合新的决策树，使参数朝着最小化损失函数的方向更新。但是，在大训练样本和高维特征的数据环境中，其效率和准确性会受到很大影响，因此，使用 histogram 算法降低内存占用率和数据分隔复杂度的 LightGBM 算法应运而生。本文选择集成 LightGBM 算法特征提取，去除初始特征集中的冗余特征，以筛选后的特征子集代替。

（二）构建舞弊识别模型

该部分分别介绍了 4 种模型的原理。逻辑回归模型主要用于解决二分类问题，在线性回归函数 $\theta^T x$ 输出实际预测值的基础上引入 *Sigmoid* 函数，寻找一个预测函数 $h_\theta(x) = g(\theta^T x)$，将实际值映射到 0，1 之间，设定一个阈值，如果 $h_\theta(x)$ 大于或等于阈值，则属于正例；如果 $h_\theta(x)$ 小于阈值，则属于负例。支持向量机模型是指通过一种非线性映射规则，将输入向量 X 映射到高维特征空间，进一步构建最优分类超平面，尽量明确正例和负例样本的分离界限，以此对数据进行分类，该模型具有很好的泛化性能。人工神经网络模型由大量神经元组成，

需要输入、输出数据进行模型训练，输入为：$\sum_{i=0}^{n-1} w_i x_i$，处理单元的输出为：$y=f(\sum_{i=0}^{n-1} w_i x_i-\theta)$，通过选择网络结构和传递函数可得到模型输出结果，然后根据实际输出值与期望值之间的误差修正权值，直到误差达到预期结果。随机森林是一种应用广泛的集成机器学习方法，可从训练集中随机、有放回地选出m个样本，共进行n次采样，生成n个训练集，然后分别训练得到n个决策树模型，每棵决策树依次选择重要特征进行分裂，最后将生成的多棵决策树按照投票法获得组成随机森林的分类结果。

（三）模型评估指标

财务舞弊识别是一个典型的二分类问题，有四种可能的分类结果，一同构成混淆矩阵。*TP*表示被正确分类为舞弊企业；*FP*表示被错误分类为舞弊企业；*FN*表示被错误分类为非舞弊企业；*TN*表示被正确分类为非舞弊企业。基于混淆矩阵采用准确率、精确率、召回率和F1值指标对分类效果进行评价，指标公式如下：

$$\text{准确率：}\ \text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{精确率：}\ \text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{召回率：}\ \text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F1分值：}\ \text{F-score}=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

准确率反映了在所有分类情况中预测正确的概率，精确率反映了舞弊样本的预测准确度。与精确率相对应的是召回率，召回率表明了真正为舞弊的样本中预测正确的情况，因此，召回率比精确率更为重要。但精确率和召回率无法同时兼顾，F1分值是精确度和召回率的调和平均数，只有当召回率和精度都处于较高水平时，F1分值才会相应提高。F1值越高，则表明模型的分类效果越好。

四、实证分析

（一）样本选择

本文以2010至2019年被沪深两市公开处罚的A股上市企业为研究样本，将国泰安数据库（CSMAR）中披露的违规类型为“虚构利润、虚列资产、虚假记载、推迟披露、重大遗漏、披露不实”的上市企业定义为舞弊样本。为保证舞弊样本不出现重复的情况，选取其中首次舞弊且被处罚的上市企业为样本，最终选取445家企业作为舞弊样本。此外，按照处于相同行业和年份、资产规模相当及未被处罚过的原则选取配对样本，最终得到890个研究样本。

（二）指标初选

已有学者的研究表明，财务指标和非财务指标都与财务舞弊有关。因此，本文在现有研究所选取的识别指标基础上，从CSMAR数据库中初步选取了反映企业偿债能力、经营能力、盈利能力等方面信息的21个财务指标，以及反映企业治理结构、股东结构和审计意见等信息的9个非财务指标，一同构成本文的初选识别指标体系，见表1。

表1 初选指标定义

指标名称			
X1流动比率	X2速动比率	X3现金比率	X4资产负债率
X5所有者权益率	X6留存收益资产比	X7应收账款与收入比	X8存货与收入比
X9流动资产周转率	X10资本密集度	X11总资产周转率	X12资产报酬率
X13总资产净利率	X14流动资产净利率	X15营业利润率	X16总资产增长率
X17可持续增长率	X18所有者权益增长率	X19每股综合收益	X20每股盈余公积
X21每股现金流量	X22股东总数	X23独董比例	X24董事长与总经理兼任情况
X25高管年薪总额	X26国有股比例	X27流通股比例	X28股权集中度
X29股权Z指数	X30审计意见类型		

（三）特征选择

基于python工具实现应用LightGBM算法识别指标的筛选，利用该算法模型对原始数据集进行训练，对得到的重要特征进行排序，本文主要提取排名靠前的10个指标作为筛选出的特征集。特征筛选结果显示，得到的特征集包含6个财务指标和4个非财务指标，分别是：X1流动比率、X6留存收益资产比、X7应收账款与收入比、X15营业利润率、X19每股综合收益、X21每股现金流量、X22股东总数、X27流通股比例、X28股权集中度及X30审

计意见类型。可以发现，这些特征分别在不同角度反映了企业的真实经营情况，指标大小与舞弊行为具有一定的相关性，这验证了应用 LightGBM 算法特征筛选的有效性。

（四）实验结果分析

经过特征筛选，最终得到初始特征和 LightGBM 筛选后的两组指标数据集，首先对初选数据集中的所有特征数据进行 z-score 标准化处理，采用十倍交叉验证的方式，运用前文介绍的 4 种机器学习技术建立模型，依托 Rapid Miner 工具实现。实现过程中，通过调整参数值，可尽量使各个模型达到最佳的分类效果，最终采用准确率、精确率、召回率和 F1 分值等指标进行评价。模型分类结果见表 2。

表2 模型分类效果对比

特征指标	识别模型	准确率	精确率	召回率	F1分值
筛选前	逻辑回归	70.27%	72.07%	65.68%	68.73%
	支持向量机	72.81%	73.55%	71.24%	72.37%
	随机森林	71.57%	72.22%	70.11%	71.15%
	人工神经网络	71.12%	71.56%	70.11%	70.83%
筛选后	逻辑回归	74.04%	75.12%	71.91%	73.48%
	支持向量机	74.83%	75.52%	73.48%	74.49%
	随机森林	73.82%	75.36%	70.79%	73.00%
	人工神经网络	72.13%	72.85%	70.56%	71.69%

由表 2 可知，基于 30 个初选特征进行分类，4 种分类器的性能与效果均能满足要求，准确率都在 70% 以上，其中支持向量机的表现最好，准确率达到 72.81%，召回率达到 71.24%，其次是随机森林、人工神经网络，逻辑回归效果表现相对较差，F1 分值性能指标的趋势与准确率一致。因此，基于初选指标集的识别，支持向量机相比其他分类器表现更好。运用 LightGBM 算法提取特征后，与使用初选特征的实验结果相比，所有分类器的准确性、召回率和 F1 分值指标都得到了提高。分类准确率最高的仍然是支持向量机，其准确率达到 74.83%，其次是逻辑回归。综合来看，LightGBM 和支持向量机相结合的分类效果最好，准确率和 F1 分值均达到最高水平。

五、结语

有效识别财务舞弊对我国经济的可持续发展具有十分重要的作用。本文基于表征企业经营情况的财务指标和非财务指标，提出应用 LightGBM 算法选择特征子集，然后基于 4 种机器学习技术构建识别模型，并对分类性能进行比较分析。结果表明：第一，通过 LightGBM 算法实现特征选择后，相较于初选特征，几乎所有的分类器性能都得到了提升，这充分验证了 LightGBM 算法在特征选择领域的有效性。第二，筛选前后，基于 2 组不同特征子集的识别，均是支持向量机模型的分类效果最好，这进一步验证了支持向量机技术的有效性。第三，验证了流动比率、营业利润率、股权集中度、审计意见等财务指标和非财务指标对舞弊识别具有重要作用。

【作者简介】于尧（1996 —），女，河北沧州人，硕士研究生，河北工业大学，研究方向为舞弊识别。