doi: 10.19920/j.cnki.jmsc.2023.10.006

基于元学习的财务舞弊识别研究①

张学勇[1]，施 懿[2]
(1. 中央财经大学金融学院，北京 100081；2. 微众银行，深圳 518063)

摘要： 本研究基于多元化数据挖掘和机器学习集成方法改进两个方面，对于如何识别财务舞弊并提高识别效率问题提供了系统性预测方法. 在多元化数据方面，不仅对传统财务因子进行了重构，而且引入公司治理层面因子并利用文本分析构建了语言类因子. 在机器学习集成方法改进方面，以 9 种不同特质的机器学习算法作为基学习器，套用元学习框架对上市公司财务舞弊进行系统性识别. 研究发现：1) 元学习框架能够显著提升舞弊样本召回率和预测精确度，提高整体学习器预测性能，并且对于大部分行业都有效果；2) 接近真实场景的滚动预测方法下，元学习框架依然能显著提高基学习器的财务舞弊识别能力；3) 公司治理因子、语言类因子对于财务舞弊识别有一定的帮助.

关键词： 财务舞弊；元学习；机器学习；文本分析
中图分类号： F832 **文献标识码：** A **文章编号：** 1007－9807(2023)10－0095－19

0 引 言

财务舞弊一直困扰着资本市场建设并愈发成为一个严重的问题[1-3]，也引起了学术界的重视[4]. 过去经验表明上市公司财务舞弊对投资者乃至整个资本市场带来了严重的负面影响. Beasley 等[5]统计发现，在初次被爆财务造假的公司在事后股价平均下跌 16.7%，并且有 47% 的财务造假公司退市. 另外，美国历史上十大破产案中有四起也与重大财务欺诈相关[6]. 财务舞弊不仅摧毁了投资者对于财务报表质量的信任，而且还严重扰乱了资本市场的正常运转. 因此，如何提高预测上市公司财务舞弊效率，这一命题在当前环境下的重要性也愈发凸显出来.

近年来，中国财务舞弊事件无论从发生频率还是涉及金额都呈现上涨趋势，愈发成为投资者、审计师、监管方所担忧的问题. 尽管财务舞弊识别需要具备丰富的经验知识以及依赖外部审计给予客观的审计意见，但从过去曝光的案例来看，传统财务舞弊识别手段由于依靠人工、效率较低，并不能较好地适应当前财务舞弊识别的需求[3]，在中国也是如此.

首先，舞弊手段逐渐变得繁复，造假模式呈现系统性、链条化，同时涉及利润表、资产负债表、现金流量表. 如新绿股份，通过伪造银行收款、虚构资金流以支撑虚增收入，再将虚增货币资金转移至在建工程、固定资产，因为固定资产折旧期限较长，折旧费用远小于虚增收入，从而使其完成对赌目标. 其次，新型金融工具、跨境业务等新形式杂糅于传统公司经营，使财务舞弊更具有隐蔽性. 例如，广东榕泰利用虚构保理业务虚增利润，宜华生活虚构境外回款虚构海外业务等. 最后，从财务舞弊披露时点来看，财务舞弊处罚具有延迟性. 根据国泰安数据统计，2006 年～2018 年上市公司财务舞弊处罚案例中，延迟 1 年～4 年占所有案例 64.15%，17.18% 的公司延迟了 5 年及以上才被

① 收稿日期：2021－01－16；修订日期：2021－12－14.
基金项目：国家哲学社会科学重大项目(19ZDA098).
作者简介：张学勇(1978—)，男，安徽庐江人，博士，教授，博士生导师. Email：zhangxueyong@cufe.edu.cn

披露其舞弊行为,财务舞弊识别的不及时,也放大了财务舞弊的恶性影响. 因此,如何提高预测上市公司财务舞弊效率,及时捕捉隐藏的财务造假信息,是解决以上问题的关键.

从过去的研究来看,大部分研究都是从财务舞弊识别因子和舞弊识别方法两个角度出发构建财务舞弊识别模型,试图以模型判断来代替人工识别. 从财务舞弊识别因子构建角度,除了传统财务指标外,研究还囊括了公司治理和文本挖掘两方面的指标,公司治理维度包括高管人员背景特征、董事会特征、股权结构等[7,8],文本挖掘维度包括财务报表附注和管理层讨论与分析(MD&A)、业绩说明会等文本的分析等[1,9-11]. 但较少有学者关注如何将财务因子、公司治理因子以及语言因子进行结合,从而提升财务舞弊识别的准确率[3]. 从舞弊识别方法角度,近年来随着计算机技术的提高以及机器学习模型的研究方法不断深入,财务舞弊识别的准确率也将逐步提高. 但目前财务舞弊识别样本存在数据不平衡的问题,一些学者仍采用 1∶1 或者固定舞弊与非舞弊样本进行模型训练,不符合真实应用场景. 并且目前选取模型的方法主要通过后验结果导向选取,而不同时期可能适用于不同的模型[12],这种主观后验选取的方法不能适应新的数据特征分布.

因此,如何提升财务舞弊识别效率并同时适应不同的数据分布特征是本文的研究目标,而本研究也将运用前沿的机器学习算法与元学习框架,将文本词性分析与金融市场相结合,探索机器学习与大数据工具在金融领域的应用. 本研究创新在于:1)结合财务因子、公司治理因子、语言因子,丰富了我国财务舞弊识别的因子库,也为财务报表文本分析应用提供了新的实证支持;2)套用元学习框架,提供了一套完整的提高财务舞弊识别准确率的方法,包括解决数据不平衡问题、因子筛选、模型调参以及模型预测,为提升财务舞弊识别预测准确率提供了可实行的客观方案.

本研究模型具体流程如图 1 所示.

1 文献综述

上市公司财务舞弊的识别不仅是投资者、审计师、监管方所担忧的问题,也是当前资本市场和国内外学术界关注的热点问题之一.

传统财务识别模型基于财务舞弊动因理论而发展起来,其中应用最为广泛的是舞弊三角理论和 GONE 理论. 舞弊三角理论将动机或压力、机会、态度或借口作为识别舞弊的重要条件,其中,实施舞弊的动机或压力是舞弊发生的首要条件. GONE 理论由 Bologna 等[13]提出,认为企业舞弊行为由 G(greed)、O(opportunity)、N(need)、E(exposure)四个影响因素驱动导致,之后他们在此基础上不断归纳,最终形成舞弊风险因子理论.

美国注册会计师协会分别在 1988 年、1997 年、2002 年发布了审计准则(SAS),明确规定了审计师审查财务舞弊的规范,因此,也开始有学者基于 SAS 所提及的风险因子构建了财务舞弊识别模型. 一方面,一些学者采用问卷、清单的形式获得 SAS 准则中提及的一些风险因子[14],但这些因子的可靠性无法确认[15]. 另一方面,SAS 几乎没有提供如何利用风险信号或者风险因子来进行判断,因此审计师还需要其他辅助工具进行判断,因此,整个审计流程是非定量且非自动化的[16].

还有一些学者用公开财务数据对财务舞弊风险进行建模. 国外较早的系统性识别财务舞弊模型是 M-score 模型,通过财务指标利用历史数据构建线性方程[17]. 之后在该基础上丰富了财务舞弊风险因子构成了 F-score 模型[18],其中 F-score 预测准确率已达 69%. 但这些模型仅涉及财务因子,事实证明公司治理、文本因子等非财务因子对于财务舞弊识别也有一定帮助[3,19-21],并且随着财务舞弊风险因子增多、模型维数上升,大量噪音、不相关的、冗余特征充斥在数据中[22],模型估计准确度也随之下降.

随着机器学习算法研究的深入和方法论的丰富,一些学者将机器学习算法应用到财务舞弊识别中[23],并且相较于传统财务识别模型有着明显的优势. 首先,机器学习擅长挖掘数据之间的非线性特征[24],并且准确率得到提升. 传统 M-score、F-score 模型仅能证明财务因子对于识别财务舞弊具有一定的解释力,而缺乏实际证据支持其预测准确效果. 而已有学者证明,机器学习准确率基本在 80% 以上[3,4]. 并且,这些不同算法中包含的激活函数、核函数形式都丰富了拟合函数以便适应预测目标的复杂性.

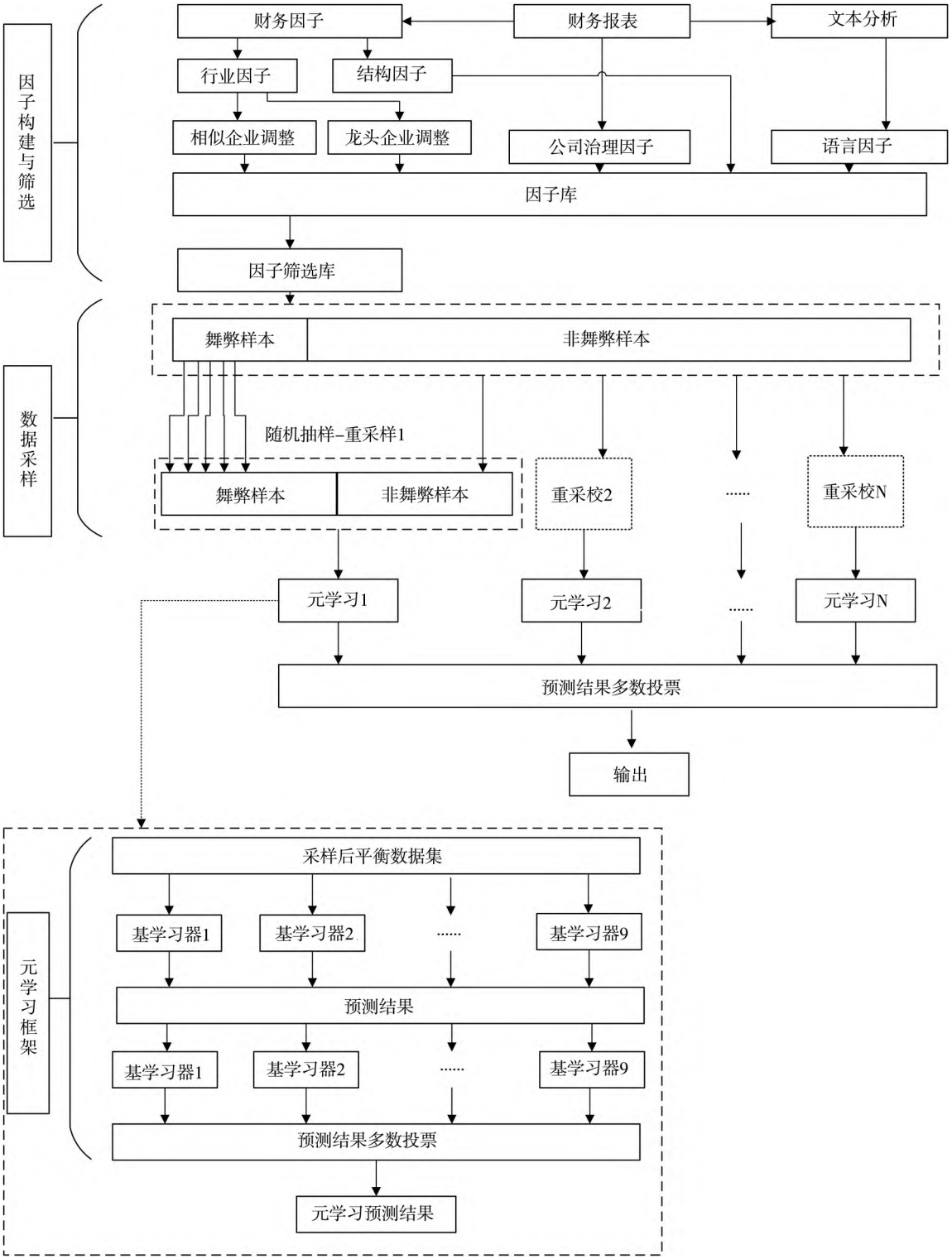

图 1　模型流程图

Fig. 1 Model flow chart

其次，机器学习可以采用集成模式而非单一模型. 传统模型通常采用单个模型进行预测，而机器学习不仅算法多样化，而且同一算法中根据不同的参数组合训练出的模型也具有多样性. 对于基础机器学习算法的优化也可以通过结合模型或者挑选适合的模型来提高预测模型的准确度[25,26]. 集成学习的思路是通过合并多个弱学习器，提升机器学习性能以获得更好的预测结果. 集

成学习分类三类: 一是用于减少偏差的 boosting, 二是用于减少方差的 bagging, 三是用于提升预测结果的 stacking. 元学习是属于 stacking, 从机器学习或者数据挖掘中得到数据, 用于提升预测结果的质量. 通常, 机器学习多种算法提供了解决问题的一系列方法, 但是并没有给出明确的答案, 在给定背景下哪些算法更适合使用, 元学习则提供了一种方式能够学习到在学习过程中哪一种算法和潜在特征能够被更有效的运用[12,27], 以解决复杂而又动态的财务舞弊识别问题, 并且随着时间推移而增强学习能力[28].

第三, 机器学习算法为如何解决数据不平衡问题提供了客观方案. 传统模型一般采用匹配样本进行解决, 但匹配近似样本的标准较为主观, 并且数据样本与真实世界不符[29]. 已有学者采用过采样的方式, 或者欠采样生成多个子数据集, 分别训练不同的模型, 再将模型进行集成输出[30], 这样的好处是利用现有全部数据, 并且避免主观挑选匹配样本, 导致训练样本与实际样本分布偏差过大, 避免出现过拟合的情况.

另外, 除了财务舞弊模型方法改进之外, 非金融变量也开始进入学术界的视野. 公司治理层面因子也有助于量化财务舞弊机会, 包括高管层持股情况、高管个人背景、股权结构等[4]. 另外, 文本分析生成的语言类因子也成为学术界金融相关研究的新工具. Humpherys 等[31] 通过词汇多样性和句法复杂性等语言因子应用于财务舞弊识别获得了接近 70% 的预测准确率. Goel 和 Uzuner[21] 则发现财务造假的公司会同时更多的使用积极和消极的词汇, 而不是采用中性词. Dong 等[10] 将文本因子分类为主体、观点、情感、情态、人称代词、写作风格、题材七个类型, 发现加入语言因子的模型平均准确率高达 82.36%, 显著优于仅采用财务比率的基准方法. Hajek 和 Henriques[3] 则得出使用否定词较少的公司存在欺诈行为的可能性较低的结论. 这也说明了数据挖掘新因子对于识别财务舞弊的重要性. 近年来, 中国学者也开始使用文本分析挖掘不同信息, 已有证据表明, 中国上市公司年报存在语调管理行为[32], 管理层报告提及的对业绩产生负面影响的内部和外部因素越多, 亏损扭转的可能性越小[33], 管理层业绩说明会上答非所问程度越高, 公司业绩则会越差[11]. 但文本分析在财务舞弊识别的问题中的应用仍相对较缺乏.

基于以上文献梳理, 本研究主要贡献在于: 1) 探索了文本分析在中国财务舞弊识别方面的应用, 构建了包含财务、公司治理、语言因子库, 扩充了财务舞弊识别因子类型; 2) 区别于仅选用一种算法的预测方式, 本研究尝试了集成学习方法在财务舞弊识别的应用, 有效提高了财务舞弊识别的准确性; 3) 相较于前人匹配近似样本时主观选取指标、或主观选取算法池的方式, 本研究套用元学习框架, 利用机器学习算法, 降低了财务舞弊识别中人工干预的程度.

2 模型流程

本研究采用的基学习器包括决策树、梯度提升树(GBDT)、K 近邻算法、Logistic 回归、朴素贝叶斯算法、随机森林算法、支持向量机、极限梯度提升树(XGBoost)、神经网络(Neural Network). 本研究主要用年报数据与季度报告数据作为基学习器的数据源, 基学习器输出的预测值作为下一层堆叠分类器的输入源, 即从底层数据输入、基学习器训练, 再到堆叠训练整个过程称为元学习.

2.1 数据处理

本研究先对全部因子作为全部解释变量进行模型训练, 以不同学习器在测试集表现的多数投票结果作为基准. 其次, 用信息增益率(*IGR*) 作为因子筛选顺序的依据, 依次去除信息增益率较小的因子, 直到基学习器多数投票结果准确率相对于基准较低为止.

信息增益(*IG*) 是评价特征对于系统的相对影响程度, 计算在该特征下的信息熵与原始信息熵的差值, 即代表该特征带给整个系统的信息增益. 利用信息增益指标有利于衡量特征对于整体系统的贡献程度, 但信息增益选择特征时容易偏向取值多的变量. 而信息增益率(*IGR*) 则克服了该问题, 信息增益率为信息增益(*IG*) 除以分裂信息度量, 即考虑了该特征数据分裂的广度和均匀程度后的信息增益. 由于在剔除冗余因子的过程中, 因子剔除顺序对于因子选择存在较大影响, 因

此,将信息增益率从小到大排序并依次剔除.

不平衡样本是财务舞弊识别模型首先需要解决的问题,因为其对于有监督机器学习任务有较大影响[34],并且强行平衡取样会导致数据样本与真实世界不符[29]. 在传统算法中,由于优化目标的设置会导致算法过多的关注多数类样本,从而使得少数类样本的分类准确度下降. 已有文献一般有两种处理方法,一是寻找匹配样本,但是对于匹配样本选取而言,匹配标准(如资产规模相近、收入相近、净利润相近等)、匹配样本比例等参数确定都存在一定主观性,并且不同参数组合对应的模型不同、预测效果也不同,存在后验偏差且不符合真实预测场景.

一般不平衡样本处理方式包括欠采样、过采样,本研究采用过采样的随机抽样算法结合集成算法对数据不平衡问题进行处理. 首先,用随机抽样过采样算法对将数据样本处理成平衡样本,接着用元学习框架对样本进行训练及样本外预测,重复随机抽样过采样算法多次,生成多个平衡样本,重复以上步骤,并对样本外预测进行预测多数投票汇总,最终得到预测结果. 这样处理的优势在于,由于随机抽样算法对于少数类样本的抽样具有随机性,训练结果可能存在随机性,生成多个平衡样本有助于降低元学习框架表现的随机性以及预测效果的随机性,且通过预测结果评价指标的波动范围能够更好的观察不同算法在元学习框架中表现的稳定性与差异性.

2.2　元学习框架

对于训练集、验证集、测试集划分,本研究分别从正常样本和舞弊样本中随机抽取 1/3 样本组成测试集,剩下的 2/3 样本作为调参数据集,用 5 折交叉验证方式进行参数网格最优搜索,用验证集表现最佳的参数作为最优模型. 本研究涉及的基学习器包括: 决策树、梯度提升树(GBDT)、K 近邻算法、Logistic 回归、朴素贝叶斯算法、随机森林算法、支持向量机、极限梯度提升树(XGBoost)、神经网络(Neural Network),每个基学习器遵循该流程进行调参.

元学习框架最好选取不同特征的基学习器[27],而最常见的分类方法包括 Logistic 回归、决策树、神经网络、支持向量机等[3],涵盖线性分类算法、概率分布算法、惰性算法、决策树、神经网络算法. 并且为了避免主观选择偏差[12],将不进行选择性集成,例如决策树、梯度提升树、等类似算法一起纳入模型中.

以下部分介绍元学习框架的预测流程. 设混合样本一共有 k 个,特征有 m 个,则元学习框架具体表示为:

1) 设原始数据为 $k \times m$ 维向量 X,对应分类标签为 $k \times 1$ 维向量 y,将数据放入基学习器中训练,优化函数为 $\min\left(\frac{1}{2k}\sum_{j=1}^{k}(y_i - f(x_i))^2\right)$,其中 f 为不同算法. 通过调参获得验证集预测准确度最高的模型,对训练集、验证集和测试集用最优模型预测得到 z.

2) 每一个分类器获得一个 $k \times 1$ 维的预测向量 z,将所有分类器预测结果堆叠形成 $k \times 9$ 维的 Z;

3) 将 Z 从基学习器输出,输入到不同的堆叠分类器中再一次通过不同机器学习算法训练,优化函数同上,对应分类标签依旧是 $k \times 1$ 维向量 y;

4) 每一个堆叠分类器输出结果为 $k \times 1$ 维向量 z',通过计算

$$y' = \begin{cases} 1, & \text{avg}(Z') > 0.5 \\ 0, & \text{avg}(Z') \leqslant 0.5 \end{cases} \tag{1}$$

获得预测值 y'.

通过元学习框架,整个流程能够自动学习各个分类器在给定情况下的优势和劣势,从底层模型的预测和分类偏差中学习,获得更大的分类能力[12,35]. 这种堆叠的成功来源于它可以利用基学习器的预测多样性,从而在元级获得更高的预测准确性,这种基学习器的再学习相较于单个分类器或者简单结合策略会更有效[6],并且在支持数据挖掘自动化方面提高学习模型的泛化能力和稳定性[36].

3　数据说明

本研究模型中的输入数据从数据来源上划分,包括年报因子和季报因子; 从类型上划分,包括财务类原始因子、财务类调整因子、公司治理因

子、语言因子.

3.1 财务舞弊样本标记

财务舞弊相较于盈余管理最大的不同在于其违法违规性质,因此,本研究采用国泰安数据库违规信息数据,以确保样本选取的客观性.

国泰安数据库中的违规信息数据中,包含了违规事件的证券代码、实际违规年份、违规类型的数据,而其中违规类型包括虚构利润、虚列资产、推迟披露、出资违规、擅自改变资金用途、内幕交易、操纵股价、违规担保等. 根据财务舞弊的性质,本研究财务舞弊样本选取的违规类型包括虚构利润、虚列资产、虚假记载(误导性陈述)、披露不实(其它),以确保财务舞弊样本类型的客观性和准确性. 财务舞弊样本标记为 1,非财务舞弊样本标记为 0.

3.2 财务因子选取以及衍生因子构建

本研究主要参照 Beneish、Cecchini 等、Abbasi 等[6,17,37] 以及文献中最常见的财务舞弊指标,从资产结构、盈利能力、现金流量、营运能力四个方面选取了 11 个公司基本面特征变量. 资产结构方面,本研究选取了资产质量指数、杠杆率[17,37]. 资产质量指数计算方式是非流动性资产/总资产,并将当期(t)除以上一期($t-1$). 若资产质量指数 > 1 说明,该公司潜在的递延成本增加,有资产夸大的可能. 杠杆率是当期总债务/总资产比上一期该值. 若杠杆率上升,债务增长相对于资产增长过快,存在财务危机可能.

盈利能力方面,本研究选取了边际利润率、净利润率、销售收入增长率[6,17,37]. 边际利润率是上一期边际利润率/当期边际利润率. 当边际利润率 > 1,说明边际利润率在恶化,说明公司可能在虚增收入而利润没有变化. 净利润率即净利润/营业利润,若企业虚增收入而没有对应的成本增加,会导致净利润上涨过快,净利润率过大. 销售收入增长率是当期销售收入/上一期销售收入,若虚增销售收入,销售收入增长率会上涨过快.

现金流量方面,本研究选取了净经营现金流与净利润的差值. 净经营现金流与净利润的差值评估应计项目对财务报表的影响[17]. 该比率如果为正,说明存在潜在收入造假的可能.

营运能力方面,本研究选取了总资产周转率、应收账款周转天数变化率、存货增长率、应收账款增长率、销售管理费用增长率[6,37]. 总资产周转率是销售收入/总资产,当虚增销售收入,会导致该值偏大. 应收账款周转天数变化率是当期(t)应收账款/销售收入比上一期($t-1$)期数据,若企业虚增收入,会导致应收账款虚增,那么应收账款周转天数也会增加. 当期存货/上一期存货. 当存货增长率值越大,说明部分销售成本转嫁到存货账面成本,虚增营业利润. 应收账款增长率是当期应收账款比上一期应收账款. 虚增收入会导致应收账款过高,应收账款增长率增加过快. 销售管理费用增长率是当期销售管理费用占销售收入比/上一期销售、管理费用占销售收入比. 若公司虚增收入,那么销售、管理费用占比会下降,销售管理费用增长率会更小.

另外,财务因子还存在一定的行业偏差和结构偏差[6]. 结构偏差是指该公司基于上一个年度或者上一个同比季度数值水平的比较,增幅或者跌幅变化. 例如该公司虽然营业利润相对于行业水平处于正常范围,但相较于上一个年度有一个极大的增幅,说明该公司可能存在营业利润调节的可能性. 而行业偏差是指该公司与行业水平的偏差. 例如该公司相较于上一年度没有异常的变动,但是其整体与行业水平相差甚远,也与同行业营业收入接近的公司水平相差甚远,这也说明了其异常的营业利润率,存在潜在财务舞弊的可能性. 因此,将行业偏差与结构偏差因素纳入模型中,也有利于财务舞弊的识别.

考虑到行业偏差和结构偏差,本研究在 11 个指标的基础上增加了衍生的行业调整因子和衍生的结构调整因子. 首先是行业调整因子. 行业调整因子分为两类,一个是样本对应同行业营业收入最接近的前 5% 家公司的对应因子平均,另一个是同行业营业收入最高的前 5% 样本因子平均,通过对比同行业接近水平以及同行业顶尖水平的公司,可以判断该公司因子值是否处于正常范围. 然后,再将原有的公司特有因子分别与同行业营业收入最接近的五个公司平均、同行业营业收入前五平均对应因子做差或者做除法. 例如净利润率的行业调整因子可以衍生出四个因子,分别是 $净利润率_{original}-净利润率_{close5\%}$ 和 $净利润率_{original}/净利润率_{close5\%}$,以及 $净利润率_{original}-净利润率_{top5\%}$

和 净利润率$_{original}$/ 净利润率$_{top5\%}$，其中下标为 original 原始数据，close5% 表示近似 5%，top5% 表示行业龙头前 5%. 其次是结构调整因子，即使用该因子对应的同比值，然后做差或者做除法. 例如净利润率的结构调整因子可以衍生出两个因子，分别是 净利润率$_t$ − 净利润率$_{t-1}$ 以及 净利润率$_t$/净利润率$_{t-1}$.

3.3　治理因子构建

学术界研究发现上市公司高管背景、高管组织结构、股本结构等与会计信息质量、财务重述行为具有一定相关性[4,38,39]. 本研究选取了高管团队背景、财报审计、股本结构、治理结构几个方面构造公司治理因子.

高管背景方面，本研究选取了高管团队女性性别占比、平均年龄、教育背景作为高管背景因子，数据来源国泰安数据库. Gao 等、Liao 等[40,41] 发现女性的存在与实施财务舞弊的低可能性相关，因此女性比例越高，财务舞弊可能性就越低. Hambrick 和 Mason[42] 认为年龄大的高管人员更加厌恶风险，因此高管平均年龄有可能与财务舞弊行为呈负相关. 卢馨等[38] 发现高管学历与财务舞弊行为严重负相关，说明学历越高财务舞弊可能性就越低.

表 1　财务原始因子

Table 1 Financial original factors

类别	因子	说明
资产结构	资产质量指数(*AQI*)	非流动性资产/总资产，并将当期(t)除以上一期($t-1$)
	杠杆率(*LEV*)	当期(t)总债务/总资产比上一期($t-1$)该值
盈利能力	边际利润率(*GMI*)	边际利润率等于销售毛利/销售收入，将上一期($t-1$)边际利润率/当期(t)边际利润率计算得到
	净利润率(*OPM*)	净利润/营业利润
	销售收入增长率(*SG*)	当期销售收入/上一期销售收入
现金流量	净经营现金流与净利润差值(*CFED*)	净经营现金流与净利润的差值
营运能力	总资产周转率(*AT*)	销售收入/总资产
	应收账款周转天数变化率(*DSIR*)	当期(t)应收账款/销售收入比上一期($t-1$)期数据
	存货增长率(*IG*)	当期存货/上一期存货
	应收账款增长率(*RG*)	应收账款增长率等于当期应收账款比上一期应收账款
	销售管理费用增长率(*SGEE*)	当期销售管理费用占销售收入比/上一期销售、管理费用占销售收入比

财报审计方面，本研究选取了是否所属四大审计事务所、当年年报审计意见、披露年报时间作为财报审计的特征因子. 四大审计事务所基于丰富经验以及出于维持其声誉的考虑，其相较于其他审计事务所的审计报告质量更高且更可信. 当年年报审计意见因子定义为标准无保留或者无保留意见加事项段为 1，保留或保留意见加事项段为 2，否定以及无法发表意见为 3. 披露年报时间定义为，该公司财务报表正式披露年报时间距离上一年 12 月 31 日之间间隔的天数. Conover 等[43] 发现财务报表披露时滞与资本市场监督程度存在相互关系，若财务报告披露越延迟，往往公司的业绩越欠佳，越有可能进行财务舞弊.

股本结构方面，本研究选取了股本结构是否变化、董事会持股数量占比、管理层持股总股数、国有股股数占比、流通股数占比、监管层持股数、股东总数作为股本结构特征. 许多学者都研究了股权结构对财务报表质量、公司治理能力的影响[39,40,44]. 总体来看，政府关联公司的治理水平较高，以及股权较分散、治理结构变动频率较低的公司财务舞弊可能性也相对较低.

治理质量方面，本研究选取了前十大股东是否存在关联、董事长与总经理是否同一人、高管团队总人数、董事人数、独立董事人数占比、监事会人数占比、董事监事及高管年薪总额作为治理质量因子. 已有研究发现董事会规模与财务舞弊有

相关性,较高的独立董事占比意味着较高的盈余信息质量[19],能够抑制财务舞弊事件的发生[14]. 另外,高管年薪与财务舞弊也有一定关系,股票薪酬的正向激励与其财务舞弊的负面效应存在矛盾[45,46],而舞弊公司往往给予高管较低的现金薪资[47].

3.4 语言因子构建

已有学者发现,年报文字所蕴含或积极或消极的信息与公司业绩、高管行为、财务报表质量有显著关系. Goel 和 Uzuner、Hajek 和 Henriques[3,21] 对美国上市公司董事会分析(MD&A)部分进行了文本分析,中国上市公司从 2005 年开始也包含了这一项内容,这一部分内容是基于管理层视角对于公司当前表现以及未来规划的文字叙述.

表 2 公司治理层面因子

Table 2 Corporate governance factors

类别	因子	说明
高管背景	女性占比(*Gender*)	女性高管比例
	平均年龄(*Age*)	求已有数据平均值
	教育背景(*Education*)	本科以下采用 0,本科为 1,硕士为 2,博士为 3;求所有高管的学历平均值
财报审计	是否四大审计事务所(*Dadtunit*)	是为 1,否为 0
	当年年报审计意见(*Audit*)	标准无保留或者无保留意见加事项段为 1,保留或保留意见加事项段为 2,否定以及无法发表意见为 3.
	披露年报时间(*Aftertime*)	距离以报表公布截止时间天数来确定
股本结构	股本结构是否变化(*isChange*)	1 = 未变化,2 = 有变化
	董事会持股数量占比(*Bddihldn*)	董事会持股占管理层持股比重
	管理层持股总股数(*Mngmhldn*)	管理层占总股数比重
	国有股股数占比(*Nshrstt*)	国有股占总股数比重
	流通股股数占比(*Nshrn*)	流通股占总股数比重
	监管层持股数(*Nshrsms*)	监管层持股数占总股数比重
	股东总数(*totalStholder*)	股东总数
治理质量	前十大股东是否存在关联(*isAssociated*)	1 = 不存在关联. 2 = 存在关联,3 = 不能确定
	董事长与总经理是否同一人(*isSame*)	1 = 同一人;2 = 不同一人
	高管团队总人数(*allnum*)	年报中披露的高级管理人员的总人数,高级管理人员含总经理,总裁,CEO,副总经理,副总裁,董秘和年报上公布的其它管理人员(包括董事中兼任的高管人员)
	董事人数(*dirnum*)	董事(含董事长)
	独立董事人数占比(*indnum*)	独立董事占董事比重
	监事会人数占比(*supnum*)	监事会人数占管理层人数比中
	董事监事及高管年薪总额(*salary*)	如未单独披露津贴的则默认为未领取任何津贴. 不包含董事、监事及高管领取的津贴.

已有文献发现财务造假者倾向于采用负面以及不确定的词汇进行表述[1],以及 MD&A 提及的负面信息越多,亏损扭转的可能性越小[33]. 并且已有证据表明,中国上市公司年报存在语调管理

行为[32]，并且管理层报告样板化对不同财务风险的企业影响存在差异[48]，因此也需要将对年报文本信息进行解析. 本研究将基于 MD&A 部分的文本进行语言分析，包括 7 个指标: 正向词(POS) 、负向词(NEG) 、情感基调(TONE) 、强烈语气(STRONG) 、模糊语气(UNCERT) 、确定性程度(CERTAIN) 、表达观点动词(REGARD) . 由于部分公司会出现更新财务报表的情况，本研究采取的语言因子采用更新后的报表 MD&A 进行文本分析. 并且中国 A 股从 2005 年才开始要求对这一部分进行强制披露，爬取文本出现较多格式错误，以及提取董事会讨论与分析文本部分规则较为混乱，因此文本因子数据从 2006 年开始.

3.5　数据说明

根据国泰安财务处罚数据统计，从图 2 可以看到每年处罚频数都在呈现上涨的趋势，并且还会出现同一年份同一家公司出现多条违规记录，并且一条记录中还出现多个财务舞弊的类型，并且平均类型数也在逐年上涨，这也说明了财务舞弊问题的严峻性. 另外，2006 年—2018 年上市公司财务舞弊处罚案例中，延迟 1 年 ~ 4 年占所有案例 64. 15%，17. 18% 的公司延迟了 5 年及以上才被披露其舞弊行为，这反映了财务舞弊具有隐蔽性，导致财务舞弊识别不及时，放大了财务舞弊的恶性影响，不过近几年财务舞弊识别延迟现象有所改善.

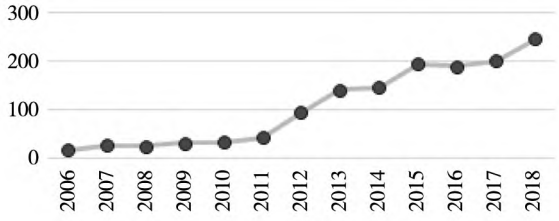

图 2　处罚记录分布

Fig. 2 The distribution of punishment record

本研究财务数据用国泰安数据库财务报表数据进行计算，文本因子从东方财富网公司公告中爬取. 先剔除了包含缺失值的样本，并对上下 1% 的异常值进行了处理，2006 年—2018 年，最终有 1 786 个舞弊样本，12 770 个非舞弊样本. 并用随机抽样算法对不平衡数据进行过采样处理，训练出模型后，再带入真实数据进行模型性能测试. 每只股票在每年所采用因子以及个数汇总如下表 3.

表 3　因子列表

Table 3 Factor list

类型	分项	总数
财务因子及其衍生因子	净经营现金流与净利润差值、存货增长率、应收账款周转天数、应收账款增长率、总资产周转率、杠杆率、净利润率、资产质量、边际利润率、销售收入增长率、销售管理费用增长率	各 28 个，共 308 个
治理因子	高管团队女性性别占比、平均年龄、教育背景、是否所属四大审计事务所、当年年报审计意见、披露年报时间、股本结构是否变化、董事会持股数量占比、管理层持股总股数、国有股股数占比、流通股数占比、监管层持股数、股东总数、前十大股东是否存在关联、董事长与总经理是否同一人、高管团队总人数、董事人数、独立董事人数占比、监事会人数占比、董事监事及高管年薪总额	各 1 个，共 20 个
语言因子	不确定因子、强烈语气因子、负向词因子、确定程度因子、正向词因子、表示观点动词因子、情感基调	各 4 个，共 28 个

4　实证结果

由于本研究为识别财务舞弊问题，相对于非舞弊样本中能够正确预测多少非舞弊样本而言，真实世界更关心预测结果为非舞弊样本中，有多少样本是真正非舞弊的公司. 同理，对于预测结果为舞弊样本中，有多少样本是真正舞弊的公司. 这样的指标称为召回率(*recall*) ，召回率更关心预测结果中的正确率.

与召回率相对应的是准确率(*precision*) ，这二者一般在同一条件下相对呈反向变化. 为了兼顾这二者指标，*F*1-*score* 作为一个综合指标可以综合考察准确率与召回率的平衡情况. 若召回率和

准确率相差较大，*F*1-*score* 将会相对较小.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2}$$

另一个模型性能指标是 *AUC* 数值，即为 ROC 曲线与坐标轴包围成的面积. ROC 曲线是按照财务舞弊识别模型预测概率进行降序排列，并累积计算模型的假正率与真正率，获得斜向上的 ROC 曲线. 若 ROC 曲线围成的面积越大，*AUC* 数值则越高，也说明模型分类效果越好.

4.1 信息增益率(*IGR*)值排序统计

每个样本在年度横截面数据中，一共有 356 个因子，最终筛选剩下 120 个因子. 以下对 *IGR* 值降序排序前 50、前 100、前 120 个因子进行统计，观察什么种类的因子对于财务舞弊识别更有效.

以上统计数据为排序靠前的因子中，来源于年报或者季报因子占其总体因子的比例，百分比含义为该属性因子占该属性原有总体因子的比例. 从上表来看，年度因子的有效性更高，说明年报数据相对于季报数据而言对于识别舞弊样本更加有效，这也比较符合直观的认识，年报相较于季报审计更为严格，因此年报因子的信息有效性程度也相较于季度因子更为高. 并且从最终挑选的因子池来看，年度因子占比高于季度因子占比 35.58% 远远高于季度因子 32.94% 的占比.

在最终因子池中被采用率最高的是原始财务因子，其占原始财务因子的 38.64%，其次是行业调整财务因子占初始因子 36.36%，紧接着是结构调整财务因子 31.82%，接着是治理因子 30%，最后是语言因子 17.86%.

从 *IGR* 值来看，单个财务因子对于识别舞弊样本起到了主要作用，并且经过行业调整及结构调整的财务因子对于整个分类系统也有较大贡献，也说明了财务因子对于识别财务舞弊的重要性. 排名前 30 财务因子对应的原始变量包括净利润率、销售收入增长率、销售管理费用增长率、边际利润率、总资产周转率，这与已有文献的观点[6,37]也是相一致的.

表 4 财务报表类型统计

Table 4 The statistics of the statements type

属性	Top 30	Top 60	筛选后	总数
季度	10.32%	26.19%	32.94%	252
年度	23.08%	32.69%	35.58%	104
总计个数	50	100	120	356

表 5 因子类型统计

Table 5 The statistics of factor type

因子	Top 50	Top 100	筛选后	总数
行业调整财务因子	14.20%	30.11%	36.36%	176
结构调整财务因子	12.50%	29.55%	31.82%	88
原始财务因子	15.91%	27.27%	38.64%	44
治理因子	25.00%	30.00%	30.00%	20
语言因子	7.14%	10.71%	17.86%	28
总计个数	50	100	120	356

对于公司治理因子而言，虽然总体因子数较少，但整体入选比例也相对较高. 因子池中最终入选的因子包括：董事会持股比例、是否为四大审计事务所、审计意见、报告披露时间、高管薪资，并且这几个因子的 *IGR* 排名也较为靠前. 从这些因子来源观察，财报审计因素对于整体公司治理因子的贡献率较高，这也侧面说明了财报审计相关因素对于判断财务舞弊也是一个重要的因素. 例如，越大的机构可能越注重声誉，其所具备能力也越专业，审计结果也越值得信赖，也就保证其审计的公司财务舞弊可能性越低. 报告披露时间距离截止时间越近，可能越欠佳，越有可能用较长的时间进行财务掩饰以及财务舞弊. 另外，董事会持股比例越高，其公司控制力越大，根据财务舞弊动因理论，其进行财务舞弊的机会也就越充分，公司财务舞弊的可能性就越大. 以及，高管薪资对于财务舞弊识别也具有一定的作用. 对于语言因子而言，在最终的因子池中只有五个因子入选，分别是第三季度确定程度因子、第一季度确定程度因子、第三季度负向因子、第二季度不确定因子以及第二季度表观点动词因子，并且分别排名第 2、第 3、第 94、第 108、第 110，这说明在第一季度、第二季度和第三季度董事会分析与讨论中，文本部分可能会释放表示确定性程度、或负向的信号，而这一信号对于识别公司舞弊将会有一定帮助.

除此之外，对比财务因子和公司治理因子，语言因子排名相对靠前的因子，全部为季报因子，而财务因子和公司治理因子中年报因子的排名靠前，这也侧面说明了对于识别财务舞弊，季报的文本信息相对于年报的文本信息更为丰富. 对于财务舞弊识别问题，投资者或者监管层更应该关注

季度报告中管理层讨论与分析部分,一些不确定的用词或者负向的表述,都有可能成为其未来业绩反转或者未来财务舞弊事件发生可能的线索和指向.

4.2　Kolmogorov-Smirnov test(K-S 检验)

为了进一步检验因子池的有效性,本研究采用 K-S 检验验证因子池的因子,在舞弊样本和非舞弊样本中存在显著差异. 按照 P 值从大到小、前 30 个因子排序结果如下表 6.

只有表示确定性程度的第三季度因子(*CERTAIN*9) 未能在 90% 的置信水平下通过检验,其余因子均在 90% 的置信水平下通过检验. 以上结果说明了大部分的因子在非舞弊和舞弊样本中存在显著的分布差异,并且在统计层面上,挑选出的因子是显著有效的,这也侧面证明了从 *IGR* 排序筛选的方法,在统计上对于因子挑选的结果也同样存在意义.

表 6　K-S 检验结果

Table 6 K-S test result

序号	因子代码	统计值	P 值	序号	因子代码	统计值	P 值
1	*CERTAIN*9	0.020 519	0.331 314	16	*AQI*(-) *_close*59	0.054 857	5.43E - 06
2	*CERTAIN*3	0.028 796	0.058 518	17	*AT*(-) *_close*59	0.055 812	3.46E - 06
3	*CFED*(-) *_close*53	0.036 658	0.006 538	18	*SG*(/) *_top*56	0.056 833	2.12E - 06
4	*Dadtunit*	0.038 747	0.003 343	19	*RG*(/) *_top*5	0.057 595	1.46E - 06
5	*DSIR*(/) *_org*9	0.040 453	0.001 88	20	*SG*(-) *_org*3	0.057 662	1.42E - 06
6	*IG*(/) *_org*9	0.041 954	0.001 11	21	*RG*(/) *_org*	0.058 654	8.66E - 07
7	*SG*(-) *_top*53	0.045 645	0.000 28	22	*SGEE*(-) *_org*3	0.058 966	7.41E - 07
8	*age*	0.045 708	0.000 273	23	*REGARD*6	0.059 051	7.10E - 07
9	*DSIR*(/) *_close*53	0.045 935	0.000 25	24	*SG*(/) *_close*59	0.059 057	7.08E - 07
10	*SG*(/) *_close*56	0.047 023	0.000 163	25	*AT*9	0.059 78	4.91E - 07
11	*Audit*	0.051 778	2.20E - 05	26	*AT*(/) *_close*53	0.060 052	4.27E - 07
12	*AT*(-) *_close*53	0.052 38	1.68E - 05	27	*SG*(/) *_top*59	0.060 462	3.46E - 07
13	*CFED*9	0.053 07	1.24E - 05	28	*DSIR*(/) *_top*59	0.060 679	3.10E - 07
14	*SG*(-) *_close*53	0.054 002	8.08E - 06	29	*NEG*3	0.061 236	2.32E - 07
15	*IG*(/) *_org*	0.054 05	7.90E - 06	30	*Bddihldn*	0.063 29	7.80E - 08

4.3　各基学习器表现

首先,在不筛选因子的情况下,各基学习器表现如表 7. 通过信息增益率进行排序并依次剔除信息增益率较小的变量,若剔除后基学习器多数投票结果的 *AUC* 值低于剔除前的 *AUC* 值,则停止剔除. 重复以上步骤获得了筛选后的因子库,并得到了各基学习器表现结果见表 8.

对比表 7 和表 8,从筛选前后预测效果来看,一方面,除了决策树和神经网络的 *AUC* 值有小幅下降外在因子筛选过后其余基学习器算法的 *AUC* 值都有了明显的提高; 从 $F1$ 值来看,除了随机森林算法外,其他算法的 $F1$ 值都有了明显提高. 另一方面,无论是因子筛选前还是筛选后,各基学习器的非舞弊精确率要明显平均高于舞弊精确率,并且因子筛选前该结论更为突出,这可能是由于一些信息增益较低的因子可能存在噪声信息,干扰了基学习器的预测.

另外,从基学习器之间预测效果对比角度(图 3) ,无论在因子筛选前还是筛选后,Logistic 回归的舞弊精确率都要明显高于非舞弊精确率,这与其他基学习器表现完全不同. 除此之外,不同基学习器的预测性能、非舞弊和舞弊精确率的权衡以及精确率和召回率的权衡上表现都不相同. 对比筛选前后,不同算法在非舞弊和舞弊精确率的权衡分布也更为分散,这也为元学习框架学习吸收不同分类器的优势提供了有利条件.

表 7 筛选前基学习器表现

Table 7 Base learner performance before selection

算法类别	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	AUC/%
Logistic 回归	0.89	12.40	98.90	22.03	50.35
朴素贝叶斯	89.81	21.38	21.19	21.29	54.36
支持向量机	98.02	10.46	6.73	8.19	50.83
k 近邻	82.35	34.05	69.18	45.64	74.92
决策树	58.71	15.20	52.79	23.60	55.80
随机森林	99.22	25.86	2.01	3.74	50.61
梯度提升树	98.69	43.64	7.30	12.51	53.01
极限梯度提升树	98.89	41.58	5.20	9.25	52.01
神经网络	93.99	17.53	8.30	11.26	52.30

注：1. 非舞弊精准率 = 预测为非舞弊样本中真实为非舞弊样本数/所有预测为非舞弊的样本数(下同).

2. 舞弊召回率 = 真实为舞弊样本中预测正确的个数/真实所为舞弊的样本数(下同).

3. 舞弊精准率 = 预测为舞弊样本中真实为舞弊样本数/所有预测为舞弊的样本数(下同).

表 8 筛选后基学习器表现

Table 8 Base learner performance after selection

算法类别	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	AUC/%
Logistic 回归	66.84	17.10	48.68	25.30	57.79
朴素贝叶斯	77.11	18.74	37.55	25.00	57.36
支持向量机	96.97	21.71	7.31	10.94	52.07
k 近邻	78.88	34.90	80.54	48.70	79.74
决策树	78.92	17.90	32.78	23.16	55.89
随机森林	99.94	0.97	0.17	0.29	50.12
梯度提升树	96.87	35.23	12.29	18.22	54.58
极限梯度提升树	94.84	34.03	19.02	24.40	56.96
神经网络	0.00	12.32	100.00	21.94	50.02

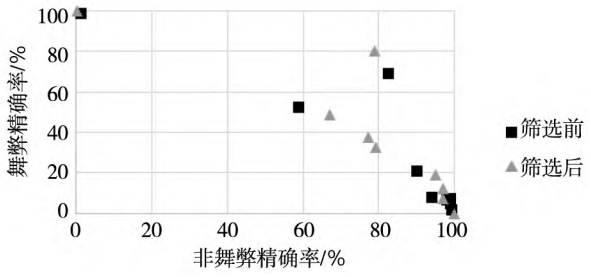

图 3 因子筛选前后基学习器预测效果对比

Fig. 3 The performance comparison between before and after selection

4.4 元学习框架下不同算法表现

将样本预测的结果进行堆叠，继续将预测结果作为解释变量，放入不同的学习器中进行进一步学习，得到单个元学习输出结果，如表 9 所示.

由元学习框架结果对比各基学习器单层学习结果对比分析发现，无论是舞弊精准率、非舞弊精准率、舞弊召回率、*AUC* 值平均水平都有显著的提高. 非舞弊精准率最高达到了 99.30%，最低精准率也达到了 98.57%，而舞弊精准率最低达到了 98.38%，最高则达到了 98.49%. 另外，舞弊召回率也均分布在 90.63% ~95.13% 的区间. 从综合评价指标 *F*1 和 *AUC* 值来看，也明显高于基学习器之前的预测水平，*F*1 值从原来的 0.29% ~48.70% 区间提高至将近 95% 以上的区间，*AUC* 指标从 50.02% ~79.74% 区间提升值 98% 以上的区间.

但通过观察发现，不同算法的结果是相同的，这是由于在堆叠各基学习器的结果后，对于同一个优化的输出结果再学习时，不同算法输出了相

似的结果，这样无法对比各算法套用元学习框架时的优劣. 因此，通过采样生成多个平衡数据集，每个数据集都堆叠出不同的输出结果，并将不同算法在各个平衡数据集的表现提取出来，便于比较不同算法在学习其他算法输出结果时的优劣. 下表为 20 个平衡数据集中不同算法的表现.

本研究展示了 20 个随机抽样数据集运用元学习训练后，以多数投票输出的结果，也做了 5 个元学习、10 个元学习、15 个元学习训练，结果也是类似的，只是标准差范围存在一定差异.

以综合指标 *F*1 和 *AUC* 作为比较标准，对比不同的算法，朴素贝叶斯和 k 近邻算法表现相对于其他算法较差，无论是 *F*1 值还是 *AUC* 值均值相对其他算法较低，并且标准差相对较高，但与其他算法相差也仅有大约 1% 的差距，也说明了元学习框架能够充分学习不同算法的预测结果，最终提高预测结果. 而相对表现较为良好的是决策树、随机森林、梯度提升树、极限梯度提升树这几种以决策树为核心的算法，其 *F*1 和 *AUC* 的均值相对较高并且标准差也较小，说明财务舞弊识别问题场景下，以非线性挖掘特征的算法能够发挥较好的优势.

表 9　元学习框架结果

Tabel 9 The performance result base on meta-learning

算法类别	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	AUC/%
Logistic 回归	99.27	94.98	98.49	96.70	98.88
朴素贝叶斯	99.21	94.56	98.38	96.43	98.79
支持向量机	99.26	94.93	98.49	96.67	98.88
k 近邻	98.57	90.63	98.60	94.45	98.59
决策树	99.30	95.13	98.38	96.72	98.84
随机森林	99.29	95.08	98.43	96.73	98.86
梯度提升树	99.26	94.93	98.49	96.67	98.88
极限梯度提升树	99.26	94.93	98.49	96.67	98.88
神经网络	98.58	93.82	98.49	96.10	98.88

表 10　20 个元学习结果

Table 10 The performance based on 20 meta-learners

算法类别	非舞弊精准率	舞弊召回率	舞弊精准率	*F*1	*AUC*
Logisitic 回归	0.992 8 ± 0.006 7	0.967 9 ± 0.017 2	0.967 8 ± 0.018 8	0.967 5 ± 0.000 9	0.989 2 ± 0.006 7
朴素贝叶斯	0.988 4 ± 0.033 1	0.963 1 ± 0.025 7	0.946 7 ± 0.041 5	0.953 7 ± 0.010 8	0.987 3 ± 0.019 0
k 近邻	0.989 6 ± 0.028 5	0.950 5 ± 0.035 4	0.966 2 ± 0.021 5	0.957 5 ± 0.008 5	0.987 5 ± 0.008 8
决策树	0.992 4 ± 0.006 8	0.967 8 ± 0.017 4	0.967 9 ± 0.019 7	0.967 5 ± 0.001 1	0.989 3 ± 0.004 4
支持向量机	0.989 6 ± 0.038 7	0.969 3 ± 0.015 9	0.959 4 ± 0.021 9	0.964 0 ± 0.003 2	0.988 0 ± 0.012 3
随机森林	0.993 0 ± 0.004 1	0.967 4 ± 0.016 7	0.968 3 ± 0.017 7	0.967 6 ± 0.000 7	0.989 1 ± 0.008 1
梯度提升树	0.992 8 ± 0.006 3	0.967 7 ± 0.016 8	0.968 7 ± 0.018 0	0.967 9 ± 0.000 7	0.989 1 ± 0.008 2
极限梯度提升树	0.992 7 ± 0.005 0	0.968 1 ± 0.017 2	0.968 0 ± 0.019 1	0.967 7 ± 0.001 0	0.989 3 ± 0.007 4
神经网络	0.992 5 ± 0.006 9	0.968 4 ± 0.017 0	0.967 0 ± 0.018 9	0.967 4 ± 0.001 1	0.988 9 ± 0.008 5

注：表格数据为均值 ± 标准差

4.5　分行业观察元学习框架性能

在预测过程中，本研究并未针对不同行业构建不同模型，即在元学习框架下并未对行业进行区分，对于所有行业均采用同一优化参数，那么元学习框架是否对于不同行业也有同样的适用度呢？

表 11 展示基于同一模型下，不同行业预测准确度是否表现良好. 一些行业由于其在 A 股上市

的规模较大，其舞弊对中国资本市场造成的影响也会更大，因此，本研究将按照证监会行业分类，重点观察上市企业较多的行业，分析其对应的舞弊识别性能如何.

表 11 分行业模型预测效果

Table 11 The performance in different industries

行业	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	AUC/%	行业样本数
制造业	99.32	95.40	98.79	97.06	99.06	9 312
信息传输、软件和信息技术服务业	99.54	97.26	98.80	98.03	99.17	959
批发和零售业	99.36	94.94	99.04	96.95	99.20	894
房地产	99.62	96.88	98.10	97.49	98.86	727
电力、热力、燃气及水的生产和供应业	99.32	95.97	99.31	97.61	99.31	513
交通运输、仓储和邮政业	99.58	94.16	97.92	96.00	98.75	381
建筑业	99.39	95.45	98.44	96.92	98.91	324
采矿	99.29	95.24	95.24	95.24	97.26	277
租赁和商务服务业	97.62	91.96	95.24	93.57	96.43	204

本研究选取了有效样本数排名前 10 的行业，可以看到 *F*1 分数几乎均在 94% 以上，*AUC* 大多在 98% 以上. 制造业作为周期性行业，业绩波动幅度较大，有较强的盈余管理以及财务造假动机，但制造业在元学习框架下，预测结果为舞弊样本中有 98.79% 为真正的舞弊样本，并且 *F*1 分数也达到了 97.06%，*AUC* 数值达 99.06%，说明元学习框架能够覆盖和有效预测制造业样本. 对于有效样本前十的行业中，舞弊样本召回率最小值也高达 91.96%，说明大部分样本中，真实为舞弊样本的公司中有高达 90% 以上的样本预测正确. 不同行业财务舞弊的手段可能是不相似的，但是粉饰业绩的指标可能是相似的，例如应收账款周转率、销售收入等. 虽然对于所有行业仅采用了一套优化参数，但是对于大部分行业的预测效果都较为稳定且良好，也说明了元学习框架在行业层面具有一定的稳健性.

4.6 滚动预测性能分析

前面的结果按照全样本进行随机抽样组成训练集、验证集和测试集，但真实世界下，当年只能知道去年全年的信息，无法预知未来的舞弊信息以及财务信息，因此，该部分用两种滚动的方式检验元学习框架的稳健性.

第一种是以五年为单位进行滚动，例如 2006 年—2010 年为一组，2011 年作为测试集，2006 年—2000 年用 5 折交叉验证进行参数调参，然后每年都以固定时间 5 年为窗口进行滚动. 第二种是时间窗口不断扩大的方式滚动，例如一开始以 2006 年—2010 年为一组，最新的 2011 年样本为测试集，2006 年—2010 年用 5 折交叉验证的方式进行参数调参; 接着以 2006 年—2011 年为一组，最新的 2012 年测试集，2006 年—2011 年用 5 折交叉验证的方式进行参数调参，以此类推.

对比表 12 和表 13，元学习框架能够较好的提高舞弊样本召回率，舞弊样本准确率也有一定的提升，说明元学习框架对于基学习器预测效果提升有较大的帮助. 虽然对比随机抽样生成的训练集、验证集、验证集而言，按照真实世界采用固定窗口滚动预测方法的 *AUC* 值有所下降，但最低也达到了 86.41%，*F*1 最低也达到了 94.59%，并且相较于单层基学习器的多数投票预测结果而言，明显有显著的提高.

表 12 基学习器固定时间 5 年窗口滚动结果

Table 12 Base learner performace based on 5-year fixed rolling window

年份	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	*AUC*/%
2011	69.92	82.38	73.13	77.48	69.05
2012	61.14	88.00	71.35	78.81	67.84
2013	67.97	81.49	71.72	76.30	67.84
2014	67.46	80.85	71.10	75.66	67.09
2015	63.01	83.19	70.05	75.99	67.34
2016	64.68	80.60	69.81	74.78	66.84
2017	57.13	86.84	68.64	76.67	66.78
2018	68.82	77.80	70.67	74.07	67.28

注: 滚动结果是样本外测试集预测结果(下同).

表 13　元学习固定时间 5 年窗口滚动结果

Table 13 Meta-learning performance based on 5-year fixed rolling window

年份	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	*AUC*/%
2011	95.33	99.33	95.78	97.51	87.08
2012	90.72	97.81	92.23	94.94	86.78
2013	92.78	98.00	93.26	95.57	86.57
2014	94.00	98.72	94.34	96.48	86.69
2015	92.87	97.55	93.25	95.35	86.51
2016	91.66	96.90	92.44	94.59	86.41
2017	91.24	97.92	91.97	94.85	86.62
2018	94.88	97.96	95.05	96.48	86.74

尝试第二种时间窗口不断扩大滚动的预测方式也得到了相同的结论(表 14、表 15).

表 14　基学习器时间窗口不断扩大滚动结果

Table 14 Base learner performance based on all-historical-data rolling window

年份	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	*AUC*/%
2011	67.54	84.99	73.54	78.78	69.44
2012	60.37	88.10	70.87	78.55	67.97
2013	59.76	87.14	70.22	77.77	67.68
2014	61.71	84.62	69.89	76.51	67.14
2015	66.17	80.77	70.32	75.18	67.52
2016	67.36	78.13	69.75	73.70	67.09
2017	64.52	80.67	69.67	74.70	67.11
2018	65.39	81.41	70.54	75.52	67.56

表 15　元学习时间窗口不断扩大滚动结果

Table 15 Meta-learning performance based on all-historical-data rolling window

年份	非舞弊精准率/%	舞弊召回率/%	舞弊精准率/%	*F*1/%	*AUC*/%
2011	94.41	98.68	95.18	96.88	87.04
2012	89.16	98.55	90.89	94.56	86.60
2013	88.45	98.43	90.37	94.23	86.61
2014	89.01	96.90	90.21	93.42	86.40
2015	92.89	97.16	93.20	95.13	86.28
2016	90.41	97.27	91.06	94.06	85.92
2017	90.97	97.23	91.82	94.44	86.31
2018	93.37	98.21	93.78	95.95	86.69

在元学习框架下，不同年份之间识别的精准率、召回率、*F*1 值都维持在 90% 以上，*AUC* 值维持在 85% 以上，并未有太大的波动，也侧面说明了元学习框架预测效果的稳健性和预测性能的优越性.

而对比固定时间窗口(表 12、表 13)的滚动以及扩大时间窗口(表 14、表 15)的滚动结果，汇总如图 4，基学习器表现以及元学习框架表现大体相差较小，但大部分的固定窗口的表现要优于扩大窗口，这一现象也说明了，可能存在舞弊企业特征与非舞弊企业特征的识别模式有较小的改变，但总体差距较小. 近年财务舞弊现象的频发并不是财务舞弊方式发生了较大的改变，而是财务舞弊可能愈发成为一种较为常见的现象，同时财务舞弊事件公布的延迟率下降、市场信息传递的效率提高，也让更多的财务舞弊事件被大众所周知.

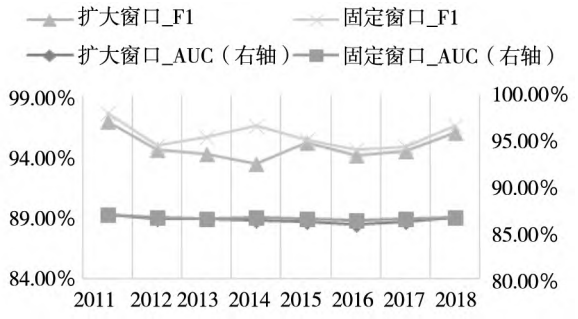

图 4　扩大窗口与固定窗口结果对比

Fig. 4 Different rolling type comparison

5　结束语

本研究对于无缺失数据的 2 437 家上市公司进行分析，涉及三大类五小类包括原始财务因子、行业调整财务因子、结构调整财务因子、公司治理因子、语言因子共计 356 个因子，并依次对因子进行筛选、异常值处理，再对数据不平衡问题处理，最后运用元学习框架对基学习器进行优化改进，以提高上市公司财务舞弊识别的准确性，并得到了如下结论:

第一，总体而言，元学习框架能够有效的提升财务舞弊识别的准确性，可以体现在提升非舞弊样本和舞弊样本精确率、舞弊样本召回率、*F*1、*AUC* 数值，这也可以说明集成算法思想的优越性. 并且元学习框架对于不同行业也同样适用，有效样本排名前十的行业召回率均在 90% 以上，说明了元学习框架的广泛适用性.

第二，在整个模型流程中，因子筛选和数据不平衡处理对于提升模型准确度也有十分重要的贡

献. 筛选因子过程采用了信息增益率加逐步筛查的方法进行,用 K-S 检验方法对因子有效性进行检验,也证明了最终筛选出的因子在舞弊样本和非舞弊样本之间存在着显著的分布差异. 另外,通过非舞弊样本精确率和舞弊样本精确率的不同算法结果分布比较,本研究发现筛选因子后不同算法区分度更高,更有利于元学习训练. 而数据不平衡处理则避免了在训练样本的过程中,过多样本一方过度学习的现象,用采样的方式解决了数据不平衡的问题,并且也提高了模型对于识别少数样本的敏感度.

第三,对于元学习框架的有效性进行了稳定性检验. 本研究模拟真实世界的信息流,采用滚动预测的方式对元学习框架进行测试,虽然结果稍逊于随机抽样生成测试集的结果,但召回率均维持在 90% 以上,*AUC* 数值均维持在 85% 以上. 并且对比不同滚动方式的预测结果,基学习器表现以及元学习框架表现大体相差较小,说明舞弊企业特征与非舞弊企业特征的识别模式仅有较小的改变.

第四,年度财务报告中提取的因子相对于季度报告因子,在识别舞弊方面更有效. 财务类因子中,结构偏差调整因子对于财务舞弊识别帮助最大,其次是行业偏差调整因子,最后是原始因子. 其次,最终有 25% 的公司治理因子被纳入最终因子库,具体包括董事会持股比例、是否为四大审计事务所、审计意见、报告披露时间、高管薪资. 语言类因子对于财务舞弊识别也有一定的帮助,其中确定性程度因子对于财务舞弊识别帮助最大,并且季报的语言类因子的有效性要高于年报的语言类因子.

总之,本研究通过元学习框架及机器学习方法,降低了以往模型的主观干预程度,有效提高了模型预测的准确性和稳定性,其中,决策树为核心的相关算法对财务舞弊识别问题上具有更大的帮助,以及语言类因子对财务舞弊识别也有一定作用.

未来工作可以从以下三个方面进行深入研究: 首先,基学习器可以进一步扩展,例如使用深度学习以进一步提高模型的预测效果,另外,对于延迟样本识别方法也有待进一步挖掘. 其次,本研究采用的字典为知网正负面评价情感词库、中国台湾大学简体中文情感词典、清华大学李军中文褒贬词典,这些字典对于解析会计文本存在一定的偏差,一些专有名词可能无法得到准确分类,未来字典的扩展和完善也有利于会计文本得到更准确的解析. 最后,本研究仅将管理层经营讨论与经营部分进行了文本词性分析,今后可以对年报其他内容进行进一步分析,也还可以对于文本分析方法进行进一步扩展,以丰富文本信息挖掘的工具并提炼更多的文本信息.

参 考 文 献:

[1] Throckmorton C S, Mayew W J, Venkatachalam M, et al. Financial fraud detection using vocal, linguistic and financial cues [J]. Decision Support Systems, 2015, 74(6): 78 – 87.

[2] Albashrawi M. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015 [J]. Journal of Data Science, 2016, 14(3): 553 – 569.

[3] Hajek P, Henriques R. Mining corporate annual reports for intelligent detection of financial statement fraud: A comparative study of machine learning methods [J]. Knowledge-Based Systems, 2017, 128(6): 139 – 152.

[4] Lin C C, Chiu A A, Huang S Y, et al. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments [J]. Knowledge-Based Systems, 2015, 89(11): 459 – 470.

[5] Beasley M S, Hermanson D R, Carcello J V, et al. Fraudulent Financial Reporting: 1998 – 2007: An Analysis of US Public Companies [R]. Durham: COSO, 2010.

[6] Abbasi A, Albrecht C, Vance A, et al. Metafraud: A meta-learning framework for detecting financial fraud [J]. MIS Quarterly, 2012, 36(4): 1293 – 1327.

[7] Chen G, Firth M, Gao D N, et al. Ownership structure, corporate governance, and fraud: Evidence from China [J]. Journal of Corporate Finance, 2006, 12(3): 424 – 448.

[8] Hasnan S, Rahman R A, Mahenthiran S. Management motive, weak governance, earnings management, and fraudulent fi-

nancial reporting: Malaysian evidence [J]. Journal of International Accounting Research, 2013, 12(1): 1 – 27.

[9] Purda L, Skillicorn D. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection [J]. Contemporary Accounting Research, 2015, 32(3): 1193 – 1223.

[10] Dong W, Liao S, Liang L. Financial Statement Fraud Detection Using Text Mining: A Systemic Functional Linguistics Theory Perspective [C]. Pacific Asia Conference on Information Systems (PACIS). Association For Information System, 2016.

[11] 卞世博，管之凡，阎志鹏. 答非所问与市场反应：基于业绩说明会的研究 [J]. 管理科学学报，2021，24(4)：109 – 126.
Bian Shibo, Guan Zhifan, Yan Zhipeng. Irrelevant answers and market reaction: Evidence from performance briefings [J]. Journal of Management Sciences in China, 2021, 24(4): 109 – 126. (in Chinese)

[12] Brazdil P, Carrier C G, Soares C, et al. Metalearning: Applications to Data Mining [M]. Berlin: Springer Science & Business Media, 2008.

[13] Bologna J, Lindquist R J, Wells J T. The Accountant's Handbook of Fraud and Commercial Crime [M]. New York: Wiley, 1993.

[14] Beasley M S. An empirical analysis of the relation between the board of director composition and financial statement fraud [J]. Accounting Review, 1996, 71(4): 443 – 465.

[15] Wilks T J, Zimbelman M F. Decomposition of fraud-risk assessments and auditors' sensitivity to fraud cues [J]. Contemporary Accounting Research, 2004, 21(3): 719 – 745.

[16] Song X P, Hu Z H, Du J G, et al. Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China [J]. Journal of Forecasting, 2014, 33(8): 611 – 626.

[17] Beneish M D. The detection of earnings manipulation [J]. Financial Analysts Journal, 1999, 55(5): 24 – 36.

[18] Bell T B, Carcello J V. A decision aid for assessing the likelihood of fraudulent financial reporting [J]. Auditing: A Journal of Practice & Theory, 2000, 19(1): 169 – 184.

[19] 胡奕明，唐松莲. 独立董事与上市公司盈余信息质量 [J]. 管理世界，2008(9)：149 – 160.
Hu Yiming, Tang Songlian. The relationship between independent directors and the quality of the information about listed companies' earnings [J]. Journal of Management World, 2008(9): 149 – 160. (in Chinese)

[20] 李　丹，宋衍蘅. 及时披露的年报信息可靠吗? [J]. 管理世界，2010(9)：129 – 137.
Li Dan, Song Yanheng. Is the timely accounting information disclosed in annual report reliable? [J]. Journal of Management World, 2010(9): 129 – 137. (in Chinese)

[21] Goel S, Uzuner O. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports [J]. Intelligent Systems in Accounting, Finance and Management, 2016, 23(3): 215 – 239.

[22] Seo J H, Choi D. Feature selection for chargeback fraud detection based on machine learning algorithms [J]. International Journal of Applied Engineering Research, 2016, 11(22): 10960 – 10966.

[23] 杨海军，太　雷. 基于模糊支持向量机的上市公司财务困境预测 [J]. 管理科学学报，2009，12(3)：102 – 110.
Yang Haijun, Tai Lei. Predicting financial distress of listed corporations based on fuzzy support vector machine [J]. Journal of Management Sciences in China, 2009, 12(3): 102 – 110. (in Chinese)

[24] Franklin J. The elements of statistical learning: Data mining, inference and prediction [J]. The Mathematical Intelligencer, 2005, 27(2): 83 – 85.

[25] Vilalta R, Drissi Y. A perspective view and survey of meta-learning [J]. Artificial Intelligence Review, 2002, 18(2): 77 – 95.

[26] Cui C, Hu M, Weir J D, et al. A recommendation system for meta-modeling: A meta-learning based approach [J]. Expert Systems with Applications, 2016, 46(3): 33 – 44.

[27] Lemke C, Budka M, Gabrys B. Metalearning: A survey of trends and technologies [J]. Artificial Intelligence Review, 2015, 44(1): 117 – 130.

[28] Matijaš M, Suykens J A K, Krajcar S. Load forecasting using a multivariate meta-learning system [J]. Expert Systems with Applications, 2013, 40(11): 4427 – 4437.

[29] Hoogs B, Kiehl T, Lacomb C, et al. A genetic algorithm approach to detecting temporal patterns indicative of financial

statement fraud [J]. Intelligent Systems in Accounting, Finance & Management: International Journal, 2007, 15(1-2): 41-56.

[30] Glancy F H, Yadav S B. A computational model for financial reporting fraud detection [J]. Decision Support Systems, 2011, 50(3): 595-601.

[31] Humpherys S L, Moffitt K C, Burns M B, et al. Identification of fraudulent financial statements using linguistic credibility analysis [J]. Decision Support Systems, 2011, 50(3): 585-594.

[32] 曾庆生, 周 波, 张 程, 等. 年报语调与内部人交易 "表里如一"还是"口是心非"? [J]. 管理世界, 2018, 34(9): 143-160.
Zeng Qingsheng, Zhou Bo, Zhang Cheng, et al. The tone of the annual report and insider trading "The same as what it says" or "Duplicity"? [J]. Journal of Management World, 2018, 34(9): 143-160. (in Chinese)

[33] 薛 爽, 肖泽忠, 潘妙丽. 管理层讨论与分析是否提供了有用信息? ——基于亏损上市公司的实证探索 [J]. 管理世界, 2010, (5): 130-140.
Xue Shuang, Xiao Zezhong, Pan Miaoli. Have the discussion and analyses of managers provided useful information? [J]. Journal of Management World, 2010, (5): 130-140. (in Chinese)

[34] Haixiang G, Yijing L, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications [J]. Expert Systems with Applications, 2017, 73(5): 220-239.

[35] Bhatt N, Thakkar A, Ganatra A, et al. Ranking of classifiers based on dataset characteristics using active meta learning [J]. International Journal of Computer Applications, 2013, 69(20): 31-36.

[36] Parmezan A R S, Lee H D, Wu F C. Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework [J]. Expert Systems with Applications, 2017, 75: 1-24.

[37] Cecchini M, Aytug H, Koehler G J, et al. Making words work: Using financial text as a predictor of financial events [J]. Decision Support Systems, 2010, 50(1): 164-175.

[38] 卢 馨, 李慧敏, 陈烁辉. 高管背景特征与财务舞弊行为的研究——基于中国上市公司的经验数据 [J]. 审计与经济研究, 2015(6): 58-68.
Lu Xin, Li Huimin, Chen Shuohui. Top managers' characteristics and financial fraud behavior: Empirical evidence from Chinese listed firms [J]. Audit and Economy Research, 2015(6): 58-68. (in Chinese)

[39] Habib A, Jiang H. Corporate governance and financial reporting quality in China: A survey of recent evidence [J]. Journal of International Accounting, Auditing and Taxation, 2015(24): 29-45.

[40] Gao Y, Kim J B, Tsang D, et al. Go before the whistle blows: An empirical analysis of director turnover and financial fraud [J]. Review of Accounting Studies, 2017, 22(1): 320-360.

[41] Liao J, Smith D, Liu X. Female CFOs and accounting fraud: Evidence from China [J]. Pacific-Basin Finance Journal, 2019, 53(2): 449-463.

[42] Hambrick D C, Mason P A. Upper echelons: The organization as a reflection of its top managers [J]. Academy of Management Review, 1984, 9(2): 193-206.

[43] Conover C M, Miller R E, Szakmary A. The timeliness of accounting disclosures in international security markets [J]. International Review of Financial Analysis, 2008, 17(5): 849-869.

[44] 陆 瑶, 张叶青, 黎 波, 等. 高管个人特征与公司业绩——基于机器学习的经验证据 [J]. 管理科学学报, 2020, 23(2): 119-139.
Lu Yao, Zhang Yeqing, Li Bo, et al. Managerial individual characteristics and corporate performance: Evidence from a machine learning approach [J]. Journal of Management Sciences in China, 2020, 23(2): 119-139. (in Chinese)

[45] Erickson M, Hanlon M, Maydew E L. Is there a link between executive equity incentives and accounting fraud? [J]. Journal of Accounting Research, 2006, 44(1): 113-143.

[46] 方军雄. 高管超额薪酬与公司治理决策 [J]. 管理世界, 2012, (11): 144-155.
Fang Junxiong. Executive excess compensation and corporate governance decision [J]. Journal of Management World, 2012, (11): 144-155. (in Chinese)

[47] Persons O S. The effects of fraud and lawsuit revelation on US executive turnover and compensation [J]. Journal of Business

Ethics, 2006, 64(4): 405 – 419.

[48] 赵子夜, 杨　庆, 杨　楠. 言多必失? 管理层报告的样板化及其经济后果 [J]. 管理科学学报, 2019, 22(3): 53 – 70.

Zhao Ziye, Yang Qing, Yang Nan. The less said the better? Economic consequences of textual similarity in management discussion and analysis [J]. Journal of Management Sciences in China, 2019, 22(3): 53 – 70. (in Chinese)

Financial fraud recognition model based on meta-learning

ZHANG Xue-yong[1], *SHI Yi*[2]

1. School of Finance, Central University of Finance and Economics, Beijing 100081, China;
2. Webank, Shenzhen 518063, China

Abstract: Based on diversified data-mining and machine learning integration, this paper provides a systematic prediction method for identifying financial fraud and improving the efficiency of financial fraud identification. In terms of diversified data-mining, not only the traditional financial factors are reconstructed, but also the corporate governance and the language factors are used. For improving the machine learning efficiency, nine machine learning algorithms with different characteristics are used as the basic learners and the meta-learning framework is applied to systematically identify the financial fraud of listed companies. This paper finds the following results. 1) Meta-learning framework can not only significantly improve the recall and precision of fraud samples and the overall prediction performance of the learner, but also is applicable for most industries. 2) Under the rolling prediction method close to the real scenario, the meta-learning can still significantly improve the financial fraud identification ability of the basic learner. 3) corporate governance factors and language factors are helpful for financial fraud identification.

Key words: financial fraud; meta-learning; machine learning; text analysis