公司财务舞弊的智能识别与模型优化策略

曾小青,唐湘勇

(长沙理工大学 经济与管理学院,湖南 长沙 410114)

摘要:公司财务舞弊行为隐蔽、手法多样,扰乱资本市场,损害投资者利益。传统方法对财务舞弊识别困难,而近年来智能财务技术发展迅速,为识别财务舞弊提供了新的途径。本文以 2010—2019 年我国 A 股上市公司为研究对象,筛选出 2 338 条舞弊数据和 4 676 条非舞弊数据进行对照,构建了基于逻辑回归、决策树、支持向量机和神经网络四类模型,并通过调参进行模型比较和优化,智能识别财务舞弊现象。研究发现,财务杠杆、流动资产周转率等十个财务指标以及大股东持股比例、审计意见两个非财务指标对模型有较好的解释力;基于多层神经网络的模型有更好的识别效果。

关键词:财务舞弊;智能财务;机器学习;舞弊识别;模型调优

[中图分类号]F406.7 [文献标识码]A [文章编号]1672-934X(2021)01-0081-12

DOI:10.16573/j.cnki.1672-934x.2021.01.010

The Strategies of Intelligent Identification and Model Optimization Respecting Corporation Financial Fraud

ZENG Xiao-qing, TANG Xiang-yong

(*School of Economics and Management, Changsha University of Science and Technology, Changsha, Hunan* 410114, *China*)

Abstract: Corporate financial fraud is a kind of concealed behavior with various means, which disrupts the capital market and damages the interests of investors. It is difficult for traditional methods to identify financial fraud, but in recent years, intelligent financial technology has developed rapidly, providing a new means for fraud identification. Taking China's A-share listed companies from 2010 to 2019 as the research object, 2,338 pieces of fraud data and 4,676 pieces of non-fraud data have been applied for comparison, and four models have been constructed based on logistic regression, decision trees, support vector machine and neural networks. Those models were compared and optimized through parameters tuning so that financial fraud could be identified intelligently. Studies found that models can be well presented through 10 financial indicators such as financial leverage, current assets turnover as well as two non-financial indicators, namely shareholding ratio of major shareholders and audit opinion, and the model based on multi-level neural network identifies financial fraud better.

Key words: financial fraud; intelligent finance; machine learning; fraud identification; model optimization

一、引言

截至 2020 年 6 月 30 日,我国深沪两市 A 股上市公司数量已达 3 936 家。然而,在资本市场不断发展壮大过程中,上市公司财务舞弊现象也屡见不鲜。公司内部人员失信、外部审

收稿日期:2020-06-16

基金项目:湖南省教育厅科学研究重点项目(19A028)

作者简介:曾小青(1975—),男,江西吉水人,副教授,博士,主要从事大数据分析、智能财务研究;

唐湘勇(1996—),女,湖南长沙人,硕士研究生,研究方向为智能财务分析。

计机构失责以及国家监管体系不完善是发生财务舞弊的主要原因[1]。近年来,瑞幸咖啡、康得新、康美药业、獐子岛等重大财务舞弊事件引起了国家和社会各界人士的广泛关注。财务舞弊行为导致财务信息失真,投资者因无法了解公司的实际情况而作出不合理的投资决策,严重损害了投资者的切身利益,同时财务舞弊行为导致资本市场资源无法得到合理配置,严重影响了资本市场的健康稳定发展。

国家证监会为加大对市场监管的力度,在2020年3月新施行的《证券法》中强调,将大幅提高财务舞弊的处罚力度,增加公司财务舞弊的成本。但财务舞弊行为隐蔽、手法多样、识别困难,使得不少上市公司为获取财务舞弊带来的巨大预期收益,铤而走险。因此,在国家出台相关政策严惩财务舞弊行为的同时,应改善财务舞弊识别手段,提高识别效果,增加舞弊成本,杜绝财务舞弊现象[2]。

二、文献综述

随着数据挖掘技术的不断发展,越来越多的学者开始研究如何利用数据挖掘技术识别财务舞弊行为。数据挖掘分类技术主要包括逻辑回归、决策树、支持向量机、贝叶斯等。逻辑回归在二分类问题中运用最广,洪文洲等人选取2003—2013年受到证监会公开处罚的44家舞弊公司和与之匹配的44家非舞弊公司作为研究样本,分别建立了采用进入法、向前逐步法和向后逐步法的Logistic回归模型,研究发现,采用进入法和向后逐步法建立的模型识别效果明显优于向前逐步法建立的模型,但受变量间多重共线性影响采用进入法的逻辑回归模型,其结果表明单个变量显著性不高[3]。考虑到模型应具备分类兼评分两个功能,学者们基于支持向量机构建财务舞弊识别模型。刘志洋等以2007—2015年首次处罚的152家制造业上市公司和152家非舞弊公司作为研究对象,研究发现,直接剔除相关性大于0.5的指标,损失了大量信息,导致Logistic回归模型识别效果不佳;采用主成分分析法降维后建立Logistic回归模型识别效果明显提高,建立具有分类和评分功能的支持向量机模型的财务舞弊识别效果最佳[4]。随着计算能力的提升和深度学习方法的改进,神经网络方法备受关注。冯炳纯将Logistic回归模型、人工神经网络、支持向量机、随机森林四种分类技术进行了对比分析[5];夏明等对数据进行主成分分析后将RBF神经网络和BP神经网络进行组合建立双隐藏层模型,与RBF神经网络和BP神经网络相比,组合模型的识别效果更佳[6]。

本文以舞弊公司所有舞弊年份中的数据作为舞弊样本,按照尽可能保留样本的完整性原则,建立了基于逻辑回归、决策树、神经网络、支持向量机的四类财务舞弊识别模型,并通过调参进行模型优化和模型比较,实现对财务舞弊现象的智能识别。

三、指标体系构建与样本选取

(一)指标体系构建

指标体系包括财务指标和非财务指标。关于财务指标选取,Persons研究发现,财务舞弊公司通常具有较高的财务杠杆率和流动资产比率、较低的资本周转率和较小的公司规模[7]。Ravisankar等对比分析了营运能力指标在舞弊样本和非舞弊样本间的差异,研究发现总资产周转率、应收账款与销售收入比有显著差异[8]。杨贵军等发现衡量偿债能力、盈利能力、现金流量、运营能力等9个财务指标对模型有较好的解释力[9]。熊方军等选取总资产周转率、资产负债率、流动负债与总负债之比等12个重要财务指标进行研究,发现通过分析财务数据波动性能有效识别上市公司的财务舞弊行为[10]。关于非财务指标选取,Rezaee研究发现公司的治理结构与财务舞弊之间存在很大的关系[11]。Gao Y等人研究发现董事会规模越大、董事会会议次数越多,则公司进行财务舞弊的可能性越小[12]。钱苹等研究发现,将审计意见、事务所规模大小等审计信息作为非财务指标加入财

务舞弊预测模型可以提高识别效果[13]。

上市公司通常采用如下手段进行财务数据造假：虚构关联方交易增加公司收入；少计费用或调整折旧少计成本；虚增银行存款、货币资本、存货等资产类科目数据；虚增利润。虚构交易时无法收到货款将导致应收账款异常增加，销售收入增长速度比成本增长速度快，因此，本文选取了应收账款与收入比、应收账款周转率、营业收入现金净含量、营业收入增长率、营业总成本增长率、营业成本率等与之相关的财务指标。少计费用和成本时，管理费用率异常故加入初始指标体系。虚增银行存款、货币资本、存货往往导致流动比率、速动比率、资产负债率、资产报酬率、总资产净利润率、流动资产周转率、流动资产净利润率、净资产收益率、存货周转率等与资产类科目相关的财务指标异常，因此，将以上财务指标加入初始指标体系。虚增利润时，资产报酬率、总资产净利润率、现金与利润总额比等反映企业盈利能力的财务指标可能存在异常，因此将其加入初始指标体系。非财务指标中，审计意见对舞弊识别作用明显，董事会会议次数及股东大会会议次数、前十大股东持股比例、董事长与总经理兼任情况等反映了公司内部治理状况。最终，本文在 CSMAR 财务指标分析中选取包括风险水平、经营能力、现金流分析、偿债能力、发展能力、盈利能力在内的 25 个财务指标，以及在外部审计、治理结构中选取 5 个非财务指标构成初始指标体系。初始特征体系具体指标如表 1 所示。

（二）样本选取

本文在国泰安数据库中选取了 2010—2019 年违规样本数据，筛选出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、披露不实等违规记录。如表 2 所示，虚假记载、推迟披露、重大遗漏等违规行为较多。在每条违规记录中按违规年度进行展开并剔除相同记录后共得到 2 668 个舞弊样本。公司若存在连续舞弊，以往学者通常选择首次舞弊年度作为违规年度，本文将公司所有舞弊年度纳入样本中，扩充了样本数据且保证了样本数据的完整性。

表 1　初始指标体系

一级指标	二级指标	三级指标及变量代码	
财务指标	风险水平	X_1	财务杠杆
		X_2	经营杠杆
		X_3	综合杠杆
	经营能力	X_4	应收账款与收入比
		X_5	应收账款周转率
		X_6	存货周转率
		X_7	流动资产周转率
		X_8	股东权益周转率
	现金流分析	X_9	营业收入现金含量
		X_{10}	营运指数
		X_{11}	资本支出与折旧摊销比
	偿债能力	X_{12}	流动比率
		X_{13}	速动比率
		X_{14}	资产负债率
	发展能力	X_{15}	固定资产增长率
		X_{16}	营业收入增长率
		X_{17}	营业总成本增长率
		X_{18}	管理费用增长率
	盈利能力	X_{19}	资产报酬率
		X_{20}	总资产净利润率（*ROA*）
		X_{21}	流动资产净利润率
		X_{22}	净资产收益率
		X_{23}	营业成本率
		X_{24}	管理费用率
		X_{25}	现金与利润总额比
非财务指标	治理结构	X_{26}	董事会会议次数
		X_{27}	股东大会召开次数
		X_{28}	十大股东持股比例
		X_{29}	董事长与总经理兼任情况
	外部审计	X_{30}	审计意见

表 2　六大违规类型数量统计

	虚构利润	虚列资产	虚假记载	推迟披露	重大遗漏	披露不实
2010 年	16	1	118	135	180	30
2011 年	26	3	162	200	276	45
2012 年	28	4	211	242	298	64
2013 年	35	4	169	251	263	78
2014 年	36	5	133	235	222	81

2015 年	52	6	152	301	260	94
2016 年	40	5	145	309	223	52
2017 年	24	3	183	306	168	24
2018 年	8	2	141	137	115	11
2019 年	1	1	42	92	81	34

利用国泰安数据库中公司研究系列中的数据收集初始指标体系中的 5 个非财务指标和 25 个财务指标数据，按图 1 所示的方法对数据进行整合，剔除空值后共得到 2 338 个舞弊样本数据。按照行业相同、会计年度相同、从未被公开处罚的原则 1∶2 选择配对样本，最终得到非舞弊样本 4 676 个。

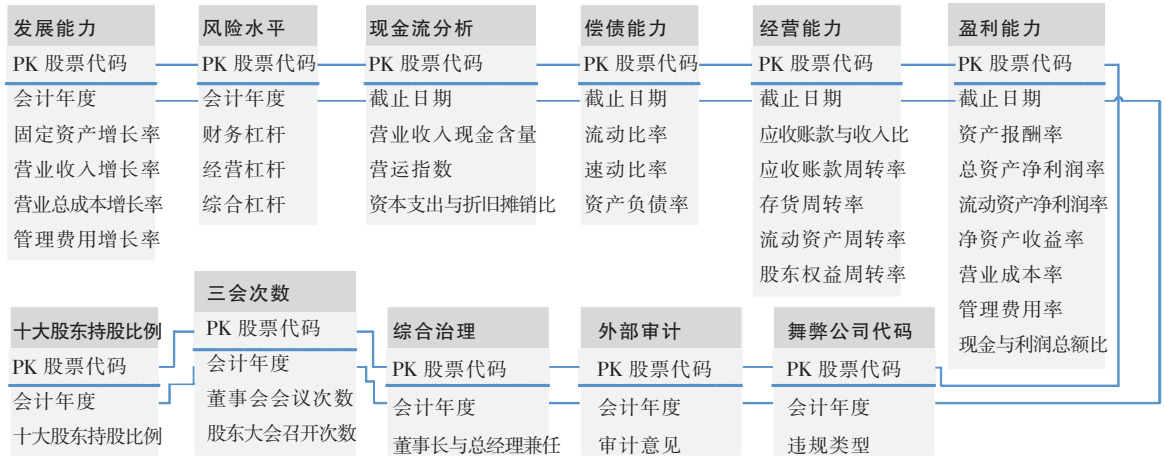

图 1　数据整合

(三)特征选择

在 SPSS 中，通过独立样本曼－惠特尼 U 检验判断 30 个变量在舞弊样本和非舞弊样本中是否存在显著差异。检验结果表明：X_6(存货周转率)、X_{16}(营业收入增长率)、X_{29}(董事长与总经理兼任情况)3 个变量不存在显著差异性，故将其剔除。

由于各指标之间具有不同的量纲，数据间差异较大。为了消除数据取值范围差异对模型效果的影响，需要对数据进行归一化处理。本文采用较为常见的“最小－最大归一化”方法。在 SPSS 中，对初始特征指标体系中样本数据进行归一化处理，处理原则如下：

$$x^* = \frac{x - \min}{\max - \min}$$

归一化处理后对 23 个财务指标进行 KMO 和巴特利特检验，结果如表 3 所示，KMO 值大于 0.5 且巴特利特球型度检验的 P 值小于 0.05，因此适合主成分分析法对数据进行降维。

在 SPSS Modeler 中按图 2 流程进行主成分分析，根据特征值大于 1.0 提取主成分，最终提取了 11 个因子，累计方差百分比为 79.110%。

表 3　KMO 和巴特利特检验

KMO 取样适切性量数		0.637
巴特利特球形度检验	近似卡方	123 878.925
	自由度	253
	显著性	0.000

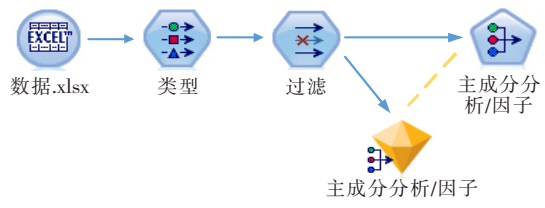

图 2　主成分分析

将 11 个因子分别命名为 Y_1，Y_2，Y_3，Y_4，Y_5，Y_6，Y_7，Y_8，Y_9，Y_{10}，Y_{11}，如表 4 所示。其中，Y_1 主要表示盈利能力指标 X_{19}，X_{20}，X_{21}，X_{22}；Y_2 主要表示偿债能力指标 X_{12}，X_{13}，X_{14}；Y_3 主

要表示发展能力指标 X_{17} 和 X_{18}；Y_4 主要表示风险水平指标 X_1 和 X_3、发展能力指标 X_{17} 和 X_{18} 以及盈利能力指标 X_{25}；Y_5 主要表示经营能力指标 X_7 和 X_8、偿债能力指标 X_{12} 和 X_{13} 以及盈利能力指标 X_{24}；Y_6 主要表示经营能力指标 X_4、盈利能力指标 X_{23} 和 X_{24}；Y_7 主要表示经营能力指标 X_5、现金流分析指标 X_9；Y_8 主要表示风险水平指标 X_2、经营能力指标 X_5、现金流分析指标 X_{11}；Y_9 主要表示现金流分析指标 X_{10} 和 X_{11}、发展能力指标 X_{15}；Y_{10} 主要表示现金流分析指标 X_{10}、发展能力指标 X_{15}；Y_{11} 主要表示风险水平指标 X_2 和发展能力指标 X_{15}。最终指标体系中含有 4 个非财务指标和 11 个因子。

表 4　主成分构成

Component	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}
X_1	−0.211	0.482	0.485	0.595	−0.088	−0.070	−0.078	0.023	−0.048	0.009	−0.030
X_2	−0.026	−0.006	0.044	0.053	0.070	0.056	0.006	−0.484	0.130	0.151	0.596
X_3	−0.214	0.455	0.494	0.607	−0.065	−0.100	−0.053	−0.031	−0.036	0.021	−0.011
X_4	−0.050	−0.079	0.049	0.062	−0.274	0.651	−0.301	−0.316	0.117	−0.033	−0.088
X_5	−0.008	0.044	0.021	0.019	0.174	−0.020	0.458	−0.405	0.349	0.182	0.099
X_7	0.045	0.442	−0.083	−0.118	0.569	0.031	0.340	−0.028	−0.012	0.003	−0.057
X_8	0.117	0.515	−0.066	−0.097	0.532	0.260	0.019	0.164	−0.038	−0.041	−0.058
X_9	−0.035	0.005	0.038	0.017	−0.180	0.429	0.588	0.232	−0.111	−0.028	−0.041
X_{10}	−0.003	−0.006	−0.016	0.005	−0.027	0.052	0.028	−0.055	−0.642	0.740	0.040
X_{11}	−0.005	−0.031	−0.001	0.006	−0.085	0.091	−0.078	0.614	0.381	0.302	0.535
X_{12}	0.156	−0.730	0.244	0.338	0.436	0.185	−0.076	0.091	−0.023	−0.001	−0.017
X_{13}	0.163	−0.724	0.247	0.343	0.444	0.185	−0.073	0.085	−0.022	−0.001	−0.019
X_{14}	−0.203	0.696	−0.077	−0.099	−0.006	0.149	−0.128	0.178	−0.008	0.009	0.043
X_{15}	0.000	−0.018	−0.026	0.017	−0.036	−0.021	−0.048	0.018	0.515	0.544	−0.566
X_{17}	0.012	0.030	−0.777	0.603	0.001	0.003	0.009	−0.006	−0.009	−0.024	0.011
X_{18}	0.027	0.030	−0.783	0.595	−0.013	0.000	0.008	0.002	0.001	−0.006	0.004
X_{19}	0.966	0.173	0.053	0.046	−0.057	0.013	−0.044	−0.007	0.004	0.004	0.007
X_{20}	0.971	0.135	0.054	0.050	−0.063	0.016	−0.052	−0.015	0.008	0.006	0.010
X_{21}	0.839	0.127	0.009	−0.007	−0.030	−0.088	0.119	−0.020	−0.008	−0.001	−0.009
X_{22}	0.898	0.251	0.070	0.060	−0.075	0.081	−0.108	−0.016	0.015	0.015	0.025
X_{23}	−0.229	0.540	−0.039	−0.053	0.262	0.422	−0.271	−0.093	0.037	0.016	0.008
X_{24}	0.026	−0.142	0.083	0.107	−0.415	0.417	0.367	0.037	−0.020	−0.072	−0.029
X_{25}	−0.075	0.215	0.321	0.401	−0.001	−0.153	0.267	0.039	0.102	−0.049	0.001

四、模型构建

在 SPSS Modeler 中对样本数据类型进行定义，按照训练集和测试集之比为 7∶3 的分配标准对样本进行分区后，分别建立基于逻辑回归、决策树、神经网络、支持向量机的财务舞弊识别模型。训练集中样本数为 4 890，测试集样本数为 2 124。建模流程如图 3 所示。

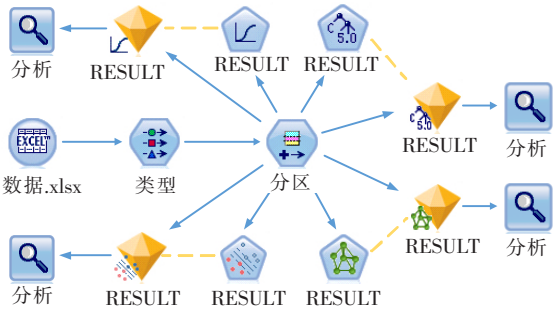

图 3　建模流程图

(一)逻辑回归

逻辑回归是二分类问题中运用最为广泛的模型。有一组自变量 $X_1, X_2, \cdots, X_n$,因变量为 Y。本文中,当 Y 表示舞弊样本时,记 $Y=1$;当 Y 表示非舞弊样本时,记 $Y=0$。逻辑回归模型中,用因变量 Y 取 0 或 1 的概率 P 来表示模型预测结果,若概率 $P(Y=1 \mid X)$ 大于 0.5 则预测结果取 1,小于 0.5 则取 0。逻辑回归模型表达式为:

$$\text{logit}(P)=ln\ (\frac{P}{1-P})=\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\cdots+\alpha_n X_n$$

等价于

$$P=P(Y=1|X)=\frac{1}{1+e^{-(\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\cdots+\alpha_n X_n)}}$$

在 SPSS Modeler 中,建立逻辑回归模型时,可以选择过程为多项式或二项式,本文选择多项式过程。多项式过程方法共有五种,分别为进入法、逐步法、前进法、后退法、后退逐步法。测试后发现,采用进入法构建的模型识别效果最优,因此本文选择使用进入法。基于 4 个非财务指标以及主成分分析法提取的 11 个主成分建立模型后,逻辑回归模型为:

$$RESULT=1/[1+exp-(-0.198\ 1*Y_1+0.514\ 9*Y_2-2.175*Y_3+1.806*Y_4-0.076\ 52*Y_5+0.196*Y_6-0.412\ 2*Y_7+0.457\ 3*Y_8-0.595\ 8*Y_9-0.442\ 5*Y_{10}+0.524\ 5*Y_{11}-0.018\ 28*X_{26}+0.093\ 11*X_{27}-0.021\ 09*X_{28}+1.821*[X_{30}=0]+0.349\ 6)]$$

表 5 参数估计

RESULT(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1.0	Intercept	0.350	0.160	4.794	1	0.029			
	Y_1	−0.198	0.031	41.229	1	0.000	0.820	0.772	0.871
	Y_2	0.515	0.041	154.318	1	0.000	1.674	1.543	1.815
	Y_3	−2.175	0.245	78.874	1	0.000	0.114	0.070	0.184
	Y_4	1.806	0.192	88.803	1	0.000	6.087	4.181	8.863
	Y_5	−0.077	0.042	3.315	1	0.069	0.926	0.853	1.006
	Y_6	0.196	0.051	14.682	1	0.000	1.217	1.100	1.345
	Y_7	−0.412	0.059	48.195	1	0.000	0.662	0.589	0.744
	Y_8	0.457	0.076	36.151	1	0.000	1.580	1.361	1.834
	Y_9	−0.596	0.171	12.159	1	0.000	0.551	0.394	0.770
	Y_{10}	−0.443	0.156	8.017	1	0.005	0.642	0.473	0.873
	Y_{11}	0.524	0.096	30.040	1	0.000	1.690	1.401	2.038
	X_{26}	−0.018	0.010	3.661	1	0.056	0.982	0.964	1.000
	X_{27}	0.093	0.021	19.562	1	0.000	1.098	1.053	1.144
	X_{28}	−0.021	0.002	88.908	1	0.000	0.979	0.975	0.983
	$[X_{30}=0.000]$	1.821	0.253	51.710	1	0.000	6.178	3.761	10.148
	$[X_{30}=1.000]$	0(b)	.	.	0	.	.	.	.

a. The reference category is: 0.0.

b. This parameter is set to zero because it is redundant.

由表 5 所示的参数估计结果可知，以 $Sig.$ 值小于 0.05 为标准，与因变量 $RESULT$ 显著相关的自变量有 Y_1，Y_2，Y_3，Y_4，Y_6，Y_7，Y_8，Y_9，Y_{10}，Y_{11}，X_{27}，X_{28}，X_{30}。逻辑回归模型的识别效果如表 6 所示，在逻辑回归模型训练集中有 3 695个样本被准确识别是否进行财务舞弊，占训练集总数的 75.56%，测试集中有 1 574 个样本被准确识别，占测试集总数的 74.11%。

表 6　逻辑回归模型识别效果

真实类别	训练集预测类别		测试集预测类别	
	0	1	0	1
0	3 119	177	1 307	73
1	1 018	576	477	267

预测结果	训练集		测试集	
	数量	占比/%	数量	占比/%
正确	3 695	75.56	1 574	74.11
错误	1 195	24.44	550	25.89
总计	4 890	100	2 124	100

（二）决策树

决策树是树形结构，由决策节点、分支和叶节点组成。决策节点表示样本属性划分；分支是对决策节点属性划分进行判断后的输出；叶节点表示经过分支判断后到达的类。决策树从顶端根节点出发，从上往下移动，每一个决策节点按照尽可能使划分后各区域样本点纯度高的原则划分属性，然后判断样本属性，最后对样本分类到达叶节点。这个自上而下进行判断输出的过程就是利用决策树进行分类的过程。

在 SPSS Modeler 中，可供选择的决策树节点有 C&R 树(R)、Quest(Q)、CHAID(C)和 C5.0 四种。C5.0 算法适合用于处理大数据，因此本文选取 C5.0 算法建立财务舞弊识别模型。数据采用主成分分析法降维后，C5.0 模型中树状图深度为 17，图 4 为部分树状图结构。

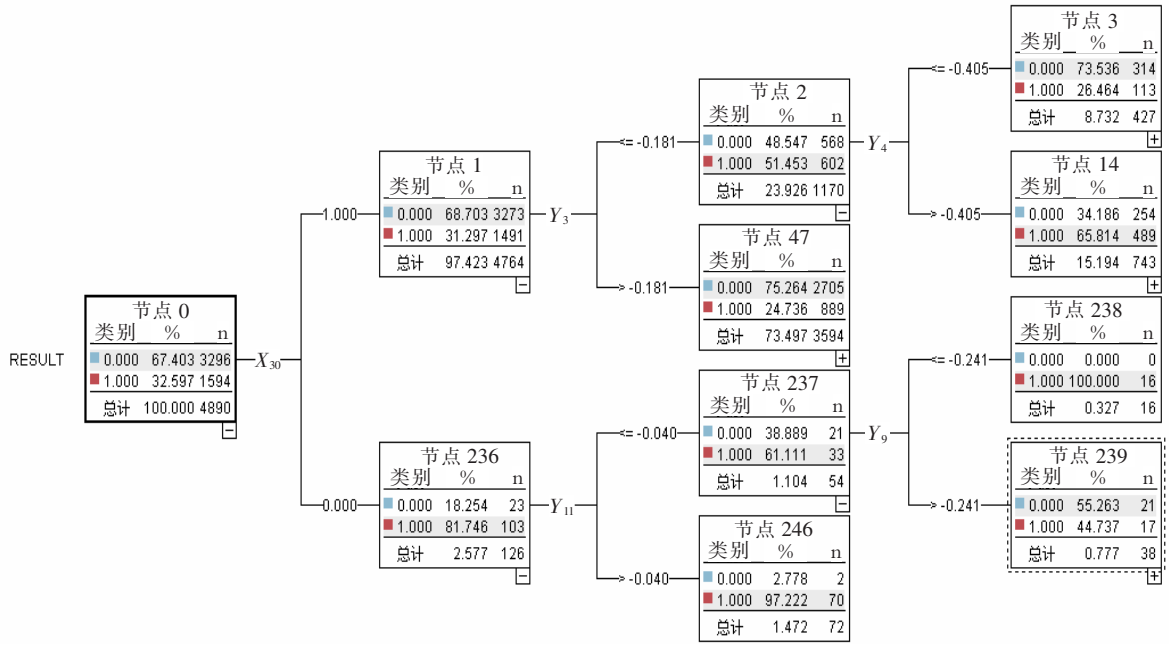

图 4　决策树树状图

决策树模型识别效果如表 7 所示，训练集中模型中有 4 410 个样本被准确识别，识别正确率为 90.18%；测试集中有 1 698 个样本被准确识别，识别正确率为 79.94%，测试集中模型识别率比训练集中的识别率降低了 10.24 个百分点。总体来说，决策树模型识别效果优于逻辑回归模型的识别效果。

决策树模型中，预测变量重要性结果如图 5 所示，排名前五的变量依次为：Y_3(0.30)，X_{30}(0.16)，Y_1(0.15)，Y_2(0.15)，Y_6(0.06)。由 Y_3

表示的指标 X_1,X_3,X_{17},X_{18}以及非财务指标 X_{30}在决策树模型中最重要。在决策树模型中,财务杠杆 X_1、综合杠杆 X_3、营业总成本增长率 X_{17}、管理费用增长率 X_{18}和审计意见 X_{30}等 5 个指标对财务舞弊识别模型有较好的解释力。

表 7　决策树模型识别效果

真实类别	训练集预测类别		测试集预测类别	
	0	1	0	1
0	3 064	232	1 186	194
1	248	1 346	232	512
预测结果	**训练集**		**测试集**	
	数量	**占比/%**	**数量**	**占比/%**
正确	4 410	90.18	1 698	79.94
错误	480	9.82	426	20.06
总计	4 890	100	2 124	100

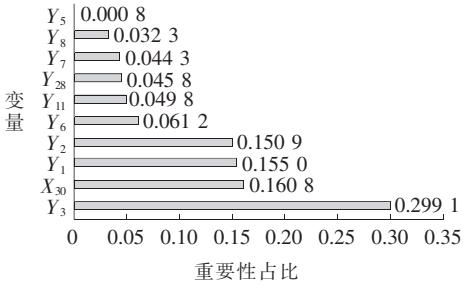

图 5　决策树模型预测变量重要性

(三)支持向量机

1.基础建模

支持向量机(SVM)将每个样本数据通过某种变换投射到空间中成为一个具体的点,使不同类别的样本点尽可能明显地区分开。支持向量机主要应用于数据样本集较小,样本维度较高的分类问题。支持向量机通过将样本的向量映射到高维空间中,寻找区分两类数据的最优超平面,使这两类数据与平面最近的点到超平面的距离最大化,与超平面的距离越大表示 SVM 的分类效果越好。支持向量机为了更好地分类,通过某种变换 $\psi(x)$,将 X 映射到高维空间 H 中,如果低维空间存在 $K(x,y)$使得:

$$K(x,y)=\psi(x)\cdot\psi(y)$$

则称 $K(x,y)$为核函数,其中 $\psi(x)\cdot\psi(y)$为 x,y 映射到空间 H 上的内积,$\psi(x)$为 $X\to H$ 的映射函数。核函数主要有四种:高斯核、线性核、多项式核、$Sigmoid$ 核。本文选择多项式核支持向量机模型,调整参数进行模型优化。SPSS Modeler 中,默认为一阶多项式,识别效果如表 8 所示。

表 8　支持向量机模型识别效果

真实类别	训练集预测类别		测试集预测类别	
	0	1	0	1
0	3 066	230	1 284	96
1	751	843	367	377
预测结果	**训练集**		**测试集**	
	数量	**占比/%**	**数量**	**占比/%**
正确	3 909	79.94	1 661	78.20
错误	981	20.06	463	21.80
总计	4 890	100	2 124	100

2.模型优化

核函数为多项式时支持向量机模型识别效果比较如表 9 所示。多项式核函数可调节的参数主要为伽马值,随着伽马值增加,模型的识别效果逐渐提高。伽马值即多项式阶数,在支持向量机模型中,随着二项式阶数增加,模型分类边界的弯曲程度会逐渐增大,分类效果越好,但是会出现模型过拟合现象,因此将导致模型在测试集中的识别效果远不及在训练集中的识别效果。且随着阶数增加,模型的运行时间也会逐渐增加,五阶多项式时运行时间为 30 分 15 秒,模型的时间成本较高。

(四)神经网络

1.基础建模

神经网络由输入层、隐藏层、输出层组成。从输入层输入变量经过神经元时会运行激活函数,对每一个输入值(x)赋予权重(w)并加上偏置(b),然后将输出结果传递给下一层的神经元。神经网络的训练过程主要包括前向传播和反向传播,前向传播是指输入变量后逐层向前

传递最后得到结果，并对比实际结果从后往前逐层逆向反馈误差，权重(w)和偏置(b)在训练过程中通过梯度下降法不断修正，然后重新进行从前往后传输，依此反复迭代直到最终预测结果与实际结果一致或者在一定的误差范围内结束训练。

在神经网络建模过程中，SPSS Modeler 中可选择的模型有径向基函数(RBF)和多层感知器(MLP)。由模型自动计算神经元数时，神经网络模型识别效果如表 10 所示。选择径向基函数时，测试集中模型识别正确率为 67.61%，识别效果不佳；选择多层感知器时，正确率为 84.09%，识别效果较好，隐藏层数和神经元数分别为 1 和 6，因此，在下文基于多层感知器调整参数进行模型优化。

表 9　多项式核支持向量机模型识别效果比较

真实类别	三阶多项式				五阶多项式			
	训练集预测类别		测试集预测类别		训练集预测类别		测试集预测类别	
	0	1	0	1	0	1	0	1
0	3 026	270	1 259	121	2 816	480	1 159	221
1	437	1 157	215	529	189	1 405	95	649
预测结果	训练集		测试集		训练集		测试集	
	数量	占比/%	数量	占比/%	数量	占比/%	数量	占比/%
正确	4 183	85.54	1 788	84.18	4 221	86.32	1 808	85.12
错误	707	14.46	336	15.82	669	13.68	316	14.88
总计	4 890	100	2 124	100	4 890	100	2 124	100

表 10　基于 RBF 和 MLP 模型的神经网络模型识别效果比较

预测结果	径向基函数				多层感知器			
	训练集		测试集		训练集		测试集	
	数量	占比/%	数量	占比/%	数量	占比/%	数量	占比/%
正确	3 433	70.20	1 436	67.61	4 185	85.58	1 786	84.09
错误	1 457	29.80	688	32.39	705	14.42	338	15.91
总计	4 890	100	4 890	100	4 890	100	2 124	100

2.模型优化

与逻辑回归、决策树、支持向量机以及采用径向基函数的神经网络模型相比，采用多层感知器的神经网络模型在测试集中的正确率最高，因此调节采用多层感知器的神经网络模型的参数对模型进行优化，以探求最佳的财务舞弊识别模型。神经网络模型可调节的参数主要有隐藏层数、隐藏层神经元数、激活函数以及梯度下降法循环次数。在 SPSS Modeler 中，神经网络模型可以调节的参数主要有隐藏层数和各隐藏层神经元数，隐藏层数最多为 2，因此在构建多层神经网络模型时存在一定局限性。构建多层神经网络时，按 n2 小于 n1 的原则调整神经元数，在测试集中的正确率最高的 9 个模型(神经网络 1－9)如表 11 所示。与自动计算神经元数得到的单层神经网络模型相比，多层神经网络模型的识别效果更佳。当 n1＝13，n2＝10 时，神经网络 7 的正确率为 86.86%，是所有神经网络模型中正确率最高的模型。

表 11　各神经网络模型测试集识别效果比较

神经网络 1(n1=6,n2=4)			神经网络 2(n1=7,n2=5)			神经网络 3(n1=8,n2=4)		
预测结果	数量	占比/%	预测结果	数量	占比/%	预测结果	数量	占比/%
正确	1 843	86.77	正确	1 822	85.78	正确	1 827	86.02
错误	281	13.23	错误	302	14.22	错误	297	13.98
总计	2 124	100	总计	2 124	100	总计	2 124	100
神经网络 4(n1=8,n2=7)			神经网络 5(n1=11,n2=6)			神经网络 6(n1=12,n2=4)		
预测结果	数量	占比/%	预测结果	数量	占比/%	预测结果	数量	占比/%
正确	1 834	86.35	正确	1 825	85.92	正确	1 818	85.59
错误	290	13.65	错误	299	14.08	错误	306	14.41
总计	2 124	100	总计	2 124	100	总计	2 124	100
神经网络 7(n1=13,n2=10)			神经网络 8(n1=13,n2=8)			神经网络 9(n1=14,n2=4)		
预测结果	数量	占比/%	预测结果	数量	占比/%	预测结果	数量	占比/%
正确	1 845	86.86	正确	1 844	86.82	正确	1 829	86.11
错误	279	13.14	错误	280	13.18	错误	295	13.89
总计	2 124	100	总计	2 124	100	总计	2 124	100

神经网络 7 如图 6 所示,神经网络包括输入层、两个隐藏层以及一个输出层。输入层有 Y_1-Y_{11},X_{26},X_{27},X_{28},X_{30} 共 15 个变量,隐藏层 1 有 13 个神经元,隐藏层 2 有 10 个神经元,输出层为 $RESULT$ 判断结果。Y_3 是最重要的变量,其次是 Y_9,Y_5,Y_6,Y_{10}。

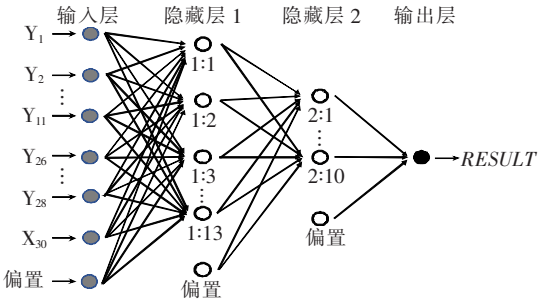

图 6　神经网络模型

神经网络 7 识别效果如表 12 所示。训练集中,$RESULT$ 分类正确的样本有 4 287 个,占训练集总数的 87.68%;测试集中,判断正确的样本数为 1 845,占测试集总数的 86.86%。神经网络模型的识别效果由于逻辑回归和决策树,识别效果较好。

表 12　神经网络 7 识别效果

真实类别	训练集预测类别		测试集预测类别	
	0	1	0	1
0	2 980	316	1 237	143
1	287	1 307	136	608
预测结果	训练集		测试集	
	数量	占比/%	数量	占比/%
正确	4 287	87.68	1 845	86.86
错误	603	12.33	279	13.14
总计	4 890	100	2 124	100

五、模型评价与比较

(一)混淆矩阵

混淆矩阵是最常用的评价二分类模型准确程度的工具,如表 13 所示。

表 13　混淆矩阵

混淆矩阵		预测值	
		1(舞弊)	0(正常)
实际值	1(舞弊)	TP	FN
	0(正常)	FP	TN

准确率（*Accuracy*），表示模型的测试组中类别判断正确的样本数占总样本数的多少。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

精确度（*Precision*），表示模型测试组预测为舞弊公司中实际舞弊公司的占比。

$$Precision=\frac{TP}{TP+FP}$$

召回度（*Recall*），也称灵敏度（*Sensitivity*），表示测试组中判断正确的舞弊公司数与实际舞弊公司数之比。

$$Recall=\frac{TP}{TP+FN}$$

F1Score 综合了精确度（*Precision*）与召回度（*Recall*）。*F1Score* 的取值范围为[0，1]，*F1Score* 越接近 1 代表模型的识别效果越好，越接近 0 代表模型的识别效果越差。

$$F1Score=\frac{2PR}{P+R}$$

（二）模型识别效果比较

基于逻辑回归、决策树、支持向量机、神经网络构建的财务舞弊识别模型效果的评价指标计算结果如表 14 所示。逻辑回归模型的召回度较低为 35.89%，决策树模型的准确率较高为 79.94%，支持向量机（五阶多项式）的召回度较高为 87.23%，多层神经网络模型 1－9 的准确率、精准度、召回度和 *F1Score* 都较高。

将以上 15 种模型的准确率、精准度、召回度、*F1Score* 进行比较可知，神经网络 7 的准确率最高为 86.86%；精准度最高的是支持向量机（三阶多项式）为 81.38%；召回度最高的是支持向量机（五阶多项式）为 87.23%，远高于其他模型；*F1Score* 最高的是神经网络 1 为 0.814 5。由此认为，基于多层神经网络模型（n1＝6，n2＝4）构建的财务舞弊识别模型有更好的识别效果。

表 14　模型评价指标计算结果

	准确率/%	精准度/%	召回度/%	*F1Score*
逻辑回归	74.11	78.53	35.89	0.492 6
决策树	79.94	72.52	68.82	0.706 2
支持向量机（一阶多项式）	78.20	79.70	50.67	0.619 6
支持向量机（三阶多项式）	84.18	81.38	71.10	0.759 0
支持向量机（五阶多项式）	85.12	74.62	87.23	0.804 2
神经网络（多层感知器）	84.09	79.42	73.66	0.764 3
神经网络 1	86.77	80.03	82.93	0.814 5
神经网络 2	85.78	80.17	78.65	0.794 0
神经网络 3	86.02	80.49	79.30	0.798 9
神经网络 4	86.35	80.19	81.05	0.806 1
神经网络 5	85.92	81.03	78.09	0.795 3
神经网络 6	85.59	79.76	78.90	0.793 2
神经网络 7	86.86	80.96	81.72	0.813 4
神经网络 8	86.82	81.02	81.45	0.812 3
神经网络 9	86.11	80.30	79.97	0.801 3

六、结论与建议

通过以上研究发现支持向量机模型和多层神经网络模型的识别效果较好。本文通过调整支持向量机多项式核函数的伽马值发现，伽马值越大，模型在训练集中识别效果越好，运行时间成本越高，模型会出现过拟合现象。通过调整隐藏层神经元数对神经网络模型进行优化得到了 9 个识别效果较好的神经网络模型。隐藏层 1 神经元数为 13，隐藏层 2 神经元数为 10 时，模型在测试集中正确率最高；隐藏层 1 的神经元数为 6，隐藏层 2 的神经元数为 4 时，模型的 *F1Score* 值最接近 1，模型整体识别效果最佳。对比支持向量机模型和神经网络模型后得出研究结论：基于多层神经网络模型建立财务舞弊识别模型的识别效果更佳。

对财务信息使用者而言,识别一个公司是否存在财务舞弊行为时,应当重点关注以下财务指标和非财务指标:财务杠杆、综合杠杆、流动资产周转率、股东权益周转率、流动比率、速动比率、资产负债率、营业总成本增长率、管理费用增长率、营业成本率、十大股东持股比例、审计意见等。上市公司主要是通过提前确认收入和虚构关联方交易来虚增收入、跨期调节费用或将费用往长期资产类科目挂账来减少负债、调减产品成本、虚增利润等方式进行财务舞弊。公司应当加强内部控制,合理增加董事会会议次数,提高公司治理效率。审计人员在对上市公司进行审计工作时,应当重点关注以上财务指标和非财务指标,并出具合理的审计意见。政府部门应当加大监管力度,同时加重对财务舞弊公司的处罚力度。

[参考文献]

[1] 郑丽萍,赵杨.上市公司财务舞弊的成因与治理研究——以瑞幸咖啡公司为例[J].管理现代化,2020(4):4-6.

[2] 黄世忠,叶钦华,徐珊.上市公司财务舞弊特征分析——基于2007年至2018年6月期间的财务舞弊样本[J].财务与会计,2019(10):24-28.

[3] 洪文洲,王旭霞,冯海旗.基于Logistic回归模型的上市公司财务报告舞弊识别研究[J].中国管理科学,2014(22):351-356.

[4] 刘志洋,韩丽荣.财务报告舞弊识别效率改善研究——基于分类技术改进和数据信息优化兼容视角[J].财经问题研究,2018(1):99-107.

[5] 冯炳纯.基于数据挖掘技术的财务舞弊识别模型构建[J].财会通讯,2019(5):93-97.

[6] 夏明,李海林,吴立源.基于神经网络组合模型的会计舞弊识别[J].统计与决策,2015(16):49-52.

[7] Persons. Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting [J]. Journal of Applied Business Research (JABR), 1995(11),38-46.

[8] Ravisankar P, Ravi V, Rao G R, et al. Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques[J].Decision Support Systems, 2011,50(2): 491-500.

[9] 杨贵军,周亚梦,孙玲莉,等.基于Benford律的Logistic模型及其在财务舞弊识别中的应用[J].统计与信息论坛,2019(8):50-56.

[10] 熊方军,张龙平.上市公司财务舞弊的风险识别与证据收集[J].经济与管理研究,2016(10):138-144.

[11] Rezaee Z. Causes, Consequences, and Deterence of Financial Statement Fraud[J]. Critical Perspectives on Accounting,2005,16(3): 277-298.

[12] 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, 2013,22(1): 320-360.

[13] 钱苹,罗玫.中国上市公司财务造假预测模型[J].会计研究,2015(7):18-25.