

面向上市企业财务报表舞弊判断的机器学习算法研究

章银平¹, 郭凤华²

(1. 安徽财贸职业学院, 安徽 合肥 230601; 2. 安徽财经大学会计学院, 安徽 蚌埠 233030)

摘要:随着我国经济的发展和社会的进步,一些上市公司为牟取私利而在财务报表上舞弊,严重扰乱了经济市场秩序,鉴于此,研究首先建立基于随机森林算法的财务舞弊判断模型,通过选择训练和测试样本,建立数学模型、设计舞弊评估方法,选择准确率、精确率、召回率和F1得分作为评价指标,对该模型判断财务舞弊的效果进行评估,并选择K邻近(KNN)、支持向量机(SVM)两种算法作为对照组。结果表明:随机森林模型准确率、精确率、召回率、F1得分和AUC值分别为0.681、0.031、0.712、0.048和0.732,均高于对照组两种算法,说明其对于舞弊样本和非舞弊样本识别能力都是最好的。

关键词:机器学习; 支持向量机; 随机森林; 财务舞弊

中图分类号: TP181 文献标识码: A 文章编号: 1673-6125 (2022) 01-0038-05

DOI:10.16856/j.cnki.52-1142/n.2022.01.020

Research on Machine Learning Algorithms for Fraud Judgment in Financial Statements of Listed Companies

ZHANG Yin-ping¹, GUO Feng-hua²

(1. Anhui Finance and Trade Vocational College, Hefei 230601, Anhui, China; 2. School of Accounting, Anhui University of Finance and Economics, Bengbu 233030, Anhui, China.)

Abstract: With the economic development and social progress of China, some listed companies cheat on financial statements in order to seek private interests, which seriously disrupts the economic and market order. In view of this, this study proposes to apply machine learning algorithm to judge the fraud of financial statements of listed enterprises. Through the establishment of machine learning-based K-proximity algorithm model, support vector machine model and random forest model, and mathematical models to evaluate the effects of the three models in judging financial fraud, the accuracy, precision, recall and F1 scores are selected as evaluation index. The results show that the recall rate of the support vector machine model is the highest, reaching 71.2%; while the random forest model has advantages in other indicators.

Key Words: Machine learning; Support vector machine; Random forest; Financial fraud

上市公司所披露的财务报表对使用者的决策起到很大影响,随着科技发展和社会进步,财务舞弊手段也更加高明,依靠传统财务报告分析手段完全判断财务舞弊很有难度^[1]。舞弊行为不断披露会使得投资者丧失对于企业的信心,继而从根本上动摇资本市场的信用根基,阻碍其发展^[2]。目前常用于财务舞弊识别的机器算法主要基于分类技术,主要有K邻近(KNN)、人工神经网络(ANN)、支持向量机(SVM)、Logistic回归和随机森林(RF)等^[3]。目前的众多研究集中于以单个分类模型对财务报告舞弊行为进行识别,关

于以集成学习算法构建财务报表舞弊识别模型的研究较少^[4]。因此研究提出建立基于集成机器学习算法的财务报表舞弊判断模型,并引入机器学习算法中的SVM模型以及KNN模型作为对照组。

1 基于随机森林算法的企业财务报表舞弊判断

1.1 基于随机森林算法的财务舞弊判断模型构建

上市企业财务报表舞弊是当前市场监管的重点内容,但是传统的舞弊监管手段非常耗时耗力,

[收稿日期] 2021-10-15

[基金项目] 2019年安徽省质量工程教师教学创新团队-会计专业创新教学团队(项目编号:2019cxtd090)

[作者简介] 章银平,男,安徽舒城人,讲师、硕士。主要研究方向:财务与会计理论和方法。

因此利用新兴技术就成为解决这个问题的主要途径^[5]。在目前的新兴技术中,计算机智能算法是应用非常广泛的一种技术,对于财务报表这种纯粹的数据信息,计算机智能算法非常擅长相关的处理工作。在目前的智能算法中,随机森林算法在样本大、数据异常时,依然能够保持较高准确性、精度和较快运算速度^[6]。因此,研究在构建上市企业财务报表舞弊判断模型时,选择随机森林算法作为模型的核心技术。基于此,本次研究选择基于机器学习算法中的随机森林算法来进行企业财务报表舞弊判断模型的构建^[9]。

随机算法作为一种集成算法,以决策树为基本单元,通过集成学习的方法完成决策树的集成。 N 棵树得到 N 个分类结果,汇总输出投票次数最多的一项^[7-8]。在随即将森林算法进行运算的过程中,其决策规则描述为:首先采用bootstrap方法抽样 N 个子样本,形成新的训练集,作为分类数,重复 M 次,得到 M 棵分类树;将输入传递到随机森林中,由决策树完成接收,每棵树预测的分类设定为独立的,并以此为标准进行投票,作为最终分类结果的判据。将 M 棵树的分类结果进行集合,通过多数投票的方式得到最终的分类结果^[10-11]。根据以上步骤,随机森林的决策规则如下图1所示。

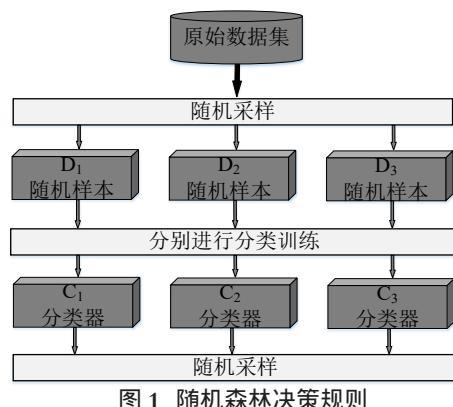


图1 随机森林决策规则

则在一个训练集 (x, y) 中,随机森林分类模型结果用下式表示:

$$F(x) = \arg \max_x \sum_{i=1}^k I(f_i(x) = y) \quad (1)$$

在式(1)中, $f_i(x)$ 表示单个决策树分类结果; y 表示输出变量,其取值范围 $[-1, 1]$,取值为 -1 表示其属于舞弊组,否则属于正常组; i 表示决策树数量,研究取 $i=100$; $I(x)$ 表示示性函数。得到一组分类模型 $h_1(x), h_2(x), \dots, h_i(x)$ 后,则用于表示样本分类正确与错误分类的平均票数之差的余量函数可表达为下式:

$$mg(x, y) = A_v I(h_k(x) = y) - \max_{j \neq y} A_v I(h_k(x) = j) \quad (2)$$

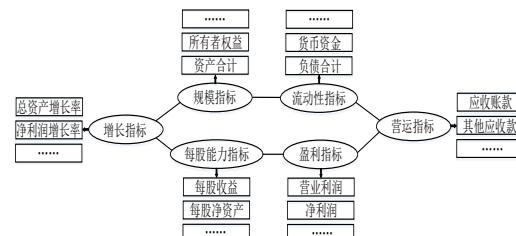
在式(2)中, $A_v(x)$ 表示取平均数。余量函数越大则分类预测效果越好、结果越可靠^[12-14]。随着决策树增加,分类模型 $h_i(x)$ 服从于强大数定律,此时泛化误差将趋于上限值,避免算法过度拟合。其中,生成决策树进行特征选择时,研究选用的最优特征选择方式是基尼指数最小化准则。对于上述样本,基尼指数表示为:

$$Gini(x, y) = 1 - \sum_{d=1}^D \left(\frac{|C_d|}{|x, y|} \right)^2 \quad (3)$$

在式(3)中, C_d 是样本中属于第 d 类的样本子集, D 是类的个数。

1.2 算法数据预处理与舞弊评估设计

现有的上市公司财务舞弊一般情况都为虚假披露盈利,但不同舞弊行为所反应在财务报告上的数据也是不尽相同的^[15]。因此研究通过对现有舞弊手段的分析,找出每种手段影响力较大的财务指标,基于此构建指标体系,从而建立判断模型。构建的财务指标体系如图2所示。



每股能力指标下的二级财务指标有基本每股收益、稀释每股收益、每股收益摊薄和每股净资产；增长指标包括基本每股收益同比增长、营业收入同比增长、营业利润同比增长、利润总额同比增长、净利润同比增长、净资产同比增长、总资产同比增长、营业利润同比增长、净利润增长率、总资产增长率和每股收益增长率这些二级指标。为保证财务指标的全面性，其来源不仅有主要财务指标，还从上市公司财务报表中覆盖了利润表、现金流量表和资产负债表。

作为典型的二分类问题，本次研究选取准确率(accuracy)、精确度(precision)、召回率(recall)、AUC以及综合计算得出的F得分(F-score)等指标来衡量模型效果。测试集样本根据真实类别以及预测类别的情况可以划分为真阳(True Positive)、假阳(False Positive)、真阴(True Negative)以及假阴(False Negative)四种情况；在真实类别中，将1记为True，0记为False；在预测类别中，将1记为Positive，0记为Negative。以此为基础建立混淆矩阵：当真实类别与预测类别都为1时，记为TP，即真阳；当真实类别取0，预测类别取1时，记为FP，即假阳；当真实类别取1，预测类别取0时，记为FN，即假阴；当真实类别与预测类别都为0时，记为TN，即真阴。则用混淆矩阵分别表示常用指标如下：

准确率即分类正确的样本在总样本之中的占比，表示如下：

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

精确率指TP样本在预测为真样本中的占比，表示如下：

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

召回率指TP样本在实际为真样本中的占比，表示如下：

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F得分表示如下：

$$F_\alpha = (1 + \alpha^2) \frac{Precision \times Recall}{\alpha^2 \times Precision + Recall} \quad (7)$$

其中， α 表示召回率的权重相对于精准率的倍数，通常取1。

研究继而引入三个指标，真阳率(True Positive Rate)又称敏感度，假阳率(False Positive Rate)表示预测为真而实际为假的样本在所有实际为假的样本中的占比，真阴率(True Negative Rate)即特异度，指预测为假并且实际为假的样本在所有实

际为假的样本中的占比。以假阳率为横轴，以真阳率为纵轴，就可以得到反馈模型效果的ROC曲线。而在ROC曲线中，曲线下方的面积定义为AUC(Area Under the Curve)，则判断模型的方法有两种，其一，曲线越靠近左上角，表示真阳率越高、假阳率越低，则模型所显示出的效果越好；其二，AUC越大，模型效果越好。真阳率、真阴率以及假阳率分别用混淆矩阵表示如下。

真阳率的表达式如下：

$$TPR = \frac{TP}{TP + FN} \quad (8)$$

真阴率的表达式如下：

$$TNR = \frac{TN}{TN + FP} \quad (9)$$

假阳率的表达式如下：

$$FPR = \frac{FP}{TN + FP} \quad (10)$$

2 模型判断效果对比与实际应用分析

为探索所建立的机器学习算法模型应用于财务舞弊判断的效果，训练和测试样本选用2010—2019年A股上市公司年度财务报表所公布的样本作为训练集，将2020年样本作为测试集，利用所构建的机器算法模型对上市公司财务舞弊概率进行判断分析。研究进行了1000次重复建模，将这些模型分别用于对测试样本的预测，并取其平均值。得到的三种模型的效果计算如表1所示。

表1 三种模型效果(阈值取值0.5)

模型效果	K邻近算法	支持向量机模型	随机森林模型
准确率	0.589	0.602	0.681
精确率	0.019	0.022	0.031
召回率	0.599	0.651	0.712
F1得分	0.038	0.044	0.048
AUC	0.690	0.721	0.732

从上表可以看出准确率、精确率、召回率、F1得分以及AUC值这5个指标的大小排序都为随机森林模型>支持向量机模型>K邻近算法，支持向量机的召回率在三个模型中是最高的，说明其对于舞弊样本的识别能力最好，而三者中对非舞弊样本识别能力最好的是随机森林模型，K邻近算法对两类样本的识别能力相对比较接近。综上，支持向量机模型和随机森林模型的效果较K邻近算法来说更优。图3绘出了不同阈值下三种模型的准确率、精确率、召回率和F1得分曲线。

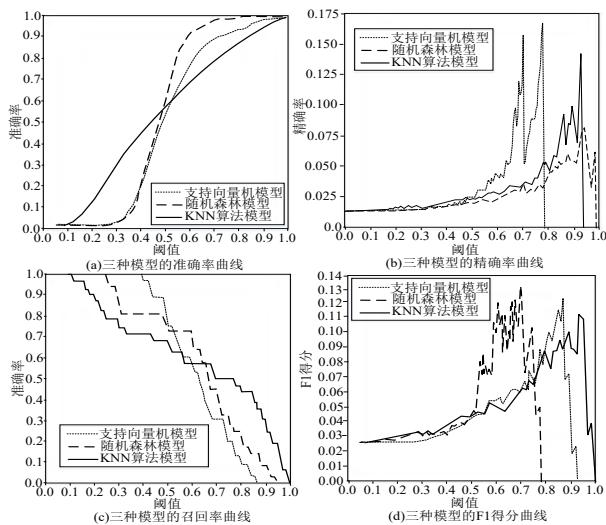


图 3 不同阈值下三种模型的准确率、精确率、召回率和 F1 得分曲线

从图 3(a)中可以看出,准确率方面,三种模型的准确率都与阈值成正比,当阈值取值范围为[0,0.5]时,KNN 算法模型的准确率高于 SVM 模型和 RF 模型;而随着阈值的增大,RF 模型的准确率在三者中达到最高;当阈值等于 0.5 时,三种模型的准确率都高于 50%。RF 模型在阈值取值范围为[0,0.2]时,几乎将所有样本错误分类,而在阈值取值范围为[0.7,1]时,将所有样本正确分类。图 3(b)显示查全率方面,三种模型的精确率都与阈值成正比,当阈值较小时,RF 模型的精确率最高,KNN 算法模型和 SVM 模型的查全率随阈值的变化相对平缓;并且,RF 模型的准确率随阈值的变化较为剧烈,且仅在阈值取值范围为[0.3,0.8]时变化。从图 3(c)、3(d)可以看出,在阈值取值[0,0.5]时,三种模型的召回率和 F1 得分差异不大。RF 模型的召回率和 F1 得分在阈值取值为[0.6,0.8]时变化剧烈,其召回率在[0.7,0.8]达到最大值,F1 得分在[0.6,0.7]达到最大值,且明显高于此时的 SVM 模型和 KNN 算法模型。



图 4 随机森林模型下舞弊概率最小的公司股票组合一年回测收益率

随机森林算法模型显示,判断概率在[0.1596, 0.1967]之间的公司可以获得大于正盈的股票收益。为进一步验证模型的效果,研究进一步选择部分公司进行时间跨度为一年的股票组合回测,通过对其收益的比较来分析模型效果,如图 4 所示。

从图中可以看出,随机森林算法模型所选出的舞弊概率最小的股票组合年化收益为 1.59%,远高于概率最大公司 21.19% 的年化损益,且高于基准收益。同时随机森林模型预测股票走势图与真实走势图十份相近,说明用随机森林算法模型选出的舞弊概率小的股票组合投资收益高、风险小,且该算法模型具有有效性和真实性。

3 结 论

研究建立基于随机森林算法的财务报表舞弊判断模型,并对其效果进行评估。结果表明 RF 模型准确率、精确率、召回率、F1 得分和 AUC 值都高于对照组两种算法,达到了 0.681、0.031、0.712、0.048 和 0.732;在阈值取值范围为[0.7,1]时,RF 模型将所有样本正确分类;当阈值较小时,RF 模型的精确率最高;RF 模型的召回率和 F1 得分在阈值取值[0.7,0.8]达到最大值,F1 得分在[0.6,0.7]达到最大值,且明显高于此时的 SVM 模型和 KNN 算法模型。以上研究说明所建立的基于随机森林算法的财务报表舞弊判断模型可用于财务舞弊的判断,证明了机器学习算法在财务舞弊判断研究领域的有效性和准确性。

参考文献:

- [1] 黄志刚,刘佳进,林朝颖.基于机器学习的上市公司财报舞弊识别前沿方法比较研究[J].系统科学与数学,2020,40(10):1882-1900.
- [2] 毛怡欣.上市公司信息披露的文字主观叙述与盈余操纵:基于机器学习法提取主观模式的研究[J].商业研究,2019(1):117-126.
- [3] 王哲,郭强,刘建国.基于机器学习和网络中心性的物流行业分析[J].物流科技,2020,43(11):7-10.
- [4] 苏颖.财务报表舞弊及治理对策研究[J].财经界,2020(12):116-117.
- [5] 诸丽行.企业内部财务报表舞弊与审计对策[J].纳税,2019,13,(7):118.
- [6] 陈健.论上市公司财务报表舞弊问题[J].福建质量管理,2019(21):48.
- [7] 张竹怡.基于舞弊三角形理论的财务报表舞弊识别的实证研究[J].纳税,2019,13(30):41-43.
- [8] 段金君.现行会计准则下上市公司财务报表舞弊及应对

- [J].中国民商,2020(3):2.
- [9]陈朝阳.银行信贷客户财务报表舞弊之甄别[J].财会学习,2020(6):120-121.
- [10]张雷鸣.财务报表中的舞弊点之分析[J].现代经济信息,2019(4):174-176.
- [11]侯民.财务报表背后的舞弊及应对策略研究[J].中国管理信息化,2019,22(10):21-22.
- [12]孙常钰.瑞幸咖啡财务舞弊及启示[J].现代商业,2020,586(33):73-74.
- [13]陈哲,许恒敏,李哲轩,等.随机生存森林:基于机器学习算法的生存分析模型[J].中华预防医学杂志,2021,55(11):104-109.
- [14]张月梅,刘媛华.基于K近邻和随机森林的情感分类研究[J].计算机与数字工程,2020,48(2):369-371.
- [15]吴芃,卢珊,杨楠.财务舞弊视角下媒体关注的公司治理角色研究[J].中央财经大学学报,2019,379(3):51-69.

[责任编辑 王建蕊]

(上接第37页)

3 结论及建议

3.1 研究结论

由层次分析法(AHP)进行长三角地区中小企业信用状况指标的权重设计,构建了长三角地区中小企业信用评价体系^[9],由计算结果可以看出,在准则层中,创新能力、区域差异、营运能力及盈利能力占比较大,四项占比达到89.59%,这说明长三角地区中小企业的创新能力、区域差异、营运能力及盈利能力对企业信用状况影响较大。从区域角度来看,长三角地区之间数字化程度、GDP等差异一方面造成了信用能力不同,另一方面也造成了信用意识的差别。在最底层中,专利级别和地区数字化程度所占比重最大,分别占比为20.62%、13.09%。这说明,专利级别较专利数量对于提升企业信用能力而言更为重要,地区数字化程度代表该地区企业数字化规模及状况,数字化程度越高则该地区中小企业利用数字化拓展业务及节省成本的比例越大,且数字化发展更利于该地区信用监管及信用体系建设。

3.2 指标体系应用建议

在该指标体系应用中,结合长三角地区中小企业现状提出以下几点信用状况提升建议:首先,协同创新提升区域数字化水平,促进中小企业信用能力及信用意识提升,大力发展战略一体化建设,提升长三角区域系统创新能力和竞争能力,让区域内中小企业更具发展潜力,提升企业信用能力,增强企业信用意识。其次,促进区域内中小企业高质量发展。提升中小企业信用能力离不开

中小企业自身高质量发展,即提升企业自身盈利及运营能力。最后,建立区域信用监管机制,打造区域信用一体化环境。为了实现长三角区域协同发展,政府应当建立适应区域发展的信用监管机制,提供良好的区域信用环境,促进区域优质的信用环境建设。

参考文献:

- [1]IMAD B H.Bayesian credit ratings:A random forest alternative approach [J].Communications in Statistics: Theory&Method,2017,46(15):7289-7300.
- [2]JADHAV S,HE H,JENKINS K.Information gain directed genetic algorithm wrapper feature selection for credit rating [J].Applied Soft Computing,2018(69):541-553.
- [3]郭林,李战江,孔文婷.基于二阶段Logistic回归的小企业信用指标判别模型[J].数学的实践与认识,2020,50(20):35-45.
- [4]丁振辉.大数据背景下的小微企业信用评级研究[J].征信,2014(11):45-49.
- [5]梁迪.小微企业信用评级指标体系研究[D].天津:天津财经大学,2015.
- [6]詹钧杰.基于AHP的国有企业信用指数研究:以山东省属上市国有企业为例[D].济南:山东大学,2020.
- [7]周桂如.基于层次分析法在酒店营运资金管理中的应用研究[J].鄂州大学学报,2020,27(6):95-97.
- [8]陈晓伟.基于改进AHP算法的移动安全支付风险评估模型[J].计算机与现代化,2020,27(11):117-121.
- [9]段翀.基于博弈论组合赋权的小微企业信用评价研究[J].征信,2019,37(9):12-17.

[责任编辑 王建蕊]