Advances in Applied Mathematics 应用数学进展, 2021, 10(12), 4227-4233
Published Online December 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1012449

Hans汉斯

基于两类Boosting的财务造假识别方法对比

范宇晨

北京信息科技大学理学院，北京

收稿日期：2021年11月13日；录用日期：2021年12月9日；发布日期：2021年12月16日

摘　要

针对财务造假问题，使用XGBoost和CatBoost两类boosting方法进行财务造假识别。对于超参数选择，通过GridSearchCV对两类boosting方法预训练，寻找最优超参数。将超参数应用于两类boosting方法上，从时间、准确度和性能的角度对比分析方法识别效果。结果表明，CatBoost方法相比于XGBoost方法，财务造假识别准确率更高，性能更优。

关键词

XGBoost，CatBoost，超参数选择，财务造假识别

The Comparison of Financial Fraud Identification Methods Based on Two Types of Boosting

Yuchen Fan

Beijing Information Science & Technology University, Beijing

Received: Nov. 13th, 2021; accepted: Dec. 9th, 2021; published: Dec. 16th, 2021

Abstract

In order to solve the financial fraud problem, XGboost and CatBoost methods are used to identify financial fraud. For the selection of super parameters, GridSearchCV is used to pre-train the two kinds of boosting methods to find the optimal super parameters. Super parameters are applied to two types of boosting methods, and the recognition effects of the methods are compared and analyzed from the perspectives of time, accuracy and performance. The results show that the CatBoost method has higher accuracy and better performance than the XGBoost method in identifying financial fraud.

文章引用: 范宇晨. 基于两类 Boosting 的财务造假识别方法对比[J]. 应用数学进展, 2021, 10(12): 4227-4233.
DOI: 10.12677/aam.2021.1012449

Keywords

XGBoost, CatBoost, Super Parameter Selection, Identification of Financial Fraud

1. 财务造假识别研究现状

财务造假识别是一个世界性的难点和热点问题。目前，财务造假识别方法多以经验法和线性模型为基础，如 Beneish 的 M-score 方法[1]、Altman 的 Z-score 方法[2]以及基于舞弊三角理论[3]的一些方法。此外，少量的研究者也在新领域对财务造假识别开展探索性研究。Qin [4]构建了向量机和逻辑回归模型，提高了财务造假识别模型的泛化和解释能力；Zheng [5]等将优化 LVQ 神经网络模型应用于财务造假识别，通过实证研究验证了模型相较于传统模型具有精准度优势。

2. boosting 决策树模型研究现状

以 boosting 决策树模型[6]为基础的 GBT 方法，在很多分类和回归挑战中都被认为是最先进、有效的。XGBoost [7]和 CatBoost [8]作为 GBT 的改进方法，在各个领域表现优异。Paleczek [9]等使用 XGBoost 方法进行糖尿病检测，结果与其他常用算法相比，XGBoost 表现出最高的性能和召回率；Thongsuwan [10]等用 XGBoost 方法改进 CNN 深度学习网络，构建了一个新的深度学习模型，简化了网络训练时误差反向传播过程并很好地处理了多数据集分类问题。针对 CatBoost，很多学者对其进行探索性研究，比如将其应用于大气 PM2.5 的预报[11]、新能源车满意度影响分析[12]等，都取得了不错的成效。

3. 方法的选择

由上述分析可知，决策树模型以其高效能、高精度的特性受到研究者广泛的青睐。结合 boosting 决策树模型的优势，本文将 XGBoost 方法和 CatBoost 方法应用于财务造假识别，通过对比分析，探索更有效、准确的财务造假识别方法。

4. boosting 决策树模型

4.1. XGBoost 模型

XGBoost 模型由 Chen [7]等提出，是对 GBT (Gradient Boosting Tree)方法的优化。GBT 模型结构图如图 1 所示，给定原始数据集 $D=\{(X_i,Y_i)\}$，其中 X_i 为特征指标数据，Y_i 为标签数据。将 X_i 输入约定的弱学习器(决策树)中，得到输出 $h_1(X_i)$，计算误差 $L(h_1(X_i),Y_i)$。对于减少误差 L，GBT 采用梯度下降的方式，训练弱学习器 $h_*(X_i)$ 以拟合 L 负梯度，将 $h_*(X_i)$ 作为增量迭代，直至 $L\left(\sum h_*(X_i),Y_i\right)$ 减小至限定的范围。训练后的弱学习器 $h_*(X_i)$，经过加权组合得到强学习器 $\sum \rho h_*(X_i)\sim Y_i$，其中 ρ 为加权系数。XGBoost 模型对 GBT 的优化主要体现在两个方面：1) XGBoost 模型在处理误差 L 上采用二阶泰勒展式，同时为了防止过拟合，对误差 L 进行了正则化处理。2) XGBoost 模型在弱学习器训练方面考虑了决策树的复杂度，不同于 ID3、CART 等决策树分裂叶子节点的方式，提出使用 Gain 增益作为分裂决策树节点的依据。XGBoost 模型相较于 GBT 模型，提高了泛化能力及稳定性。

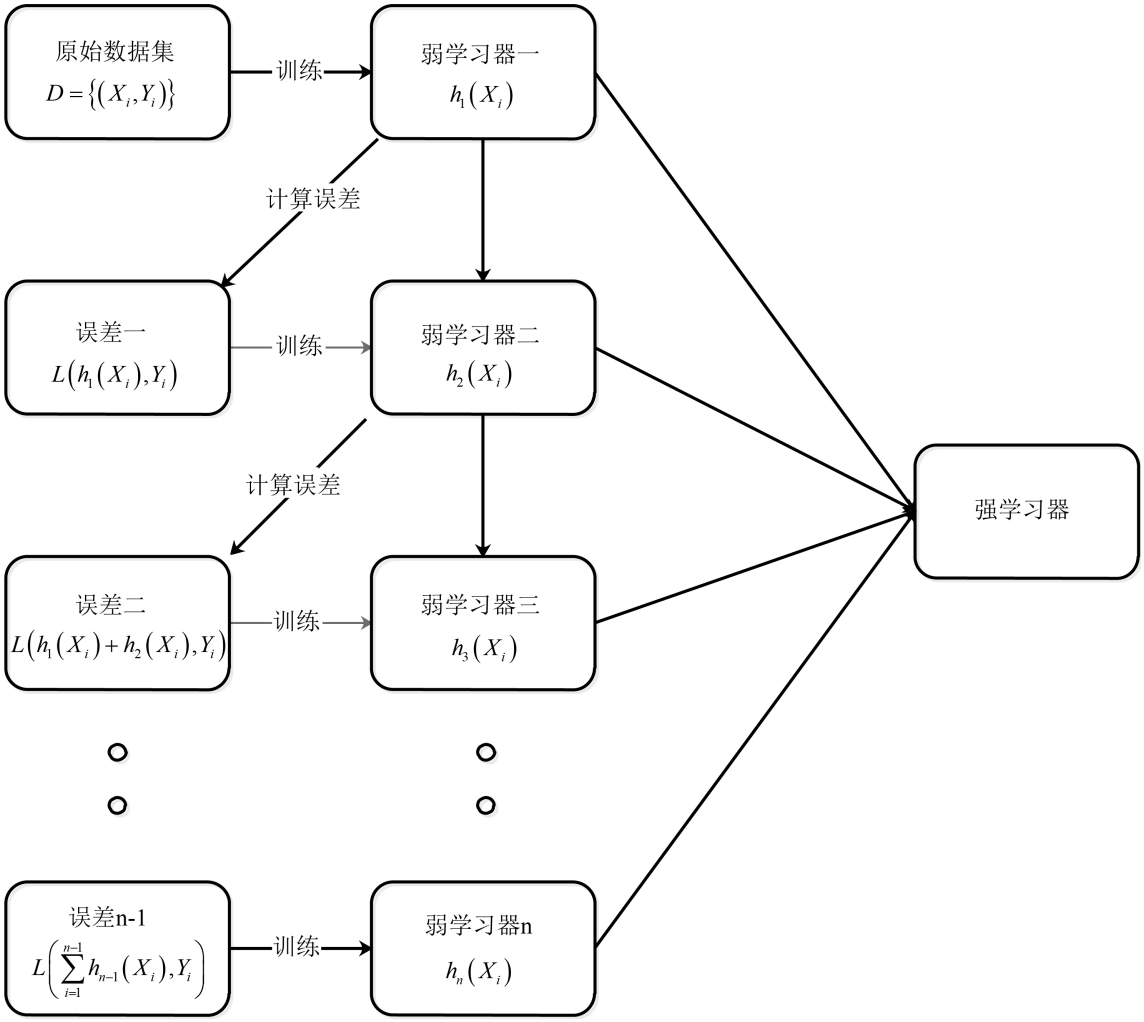

Figure 1. Diagrammatic figure of GBT
图 1. GBT 模型图

4.2. CatBoost 模型

CatBoost 模型最早由 Yandex 公司提出，是继 XGBoost 之后一类新型 boosting 方法。CatBoost 主要改进了 GBT 中存在的预测偏移问题[8]。GBT 在每次训练弱学习器过程中，由于使用的是同一个的数据集，对误差 L 负梯度的估计并非是无偏估计，因此 CatBoost 提出 Ordered boosting 排序算法，在训练数据前，打乱数据排列同时交叉训练弱学习器，以得到误差 L 负梯度的无偏估计，减少弱学习器拟合的偏差，提高模型的泛化能力。

5. 财务造假识别

本文的模型主要针对美国上市公司财务造假案例进行识别分析。

5.1. 财务数据集来源

本文的训练样本来源于 Bao [13]等在 github 上公布的数据。该数据中财务欺诈样本来源于 AAER 数据库，由加州大学伯克利分校编制管理，相应的财务指标数据来源于 COMPUSTAT 基本年度数据库。数

据文件“uscecchini28.csv”中包含 1990~2014 年美国上市公司财务造假标签数据和财务指标数据，总计 146,045 条记录数据。其中，造假标签数据“misstate”由布尔型变量(0, 1)描述，财务指标由 28 个基本财务指标和 14 个财务比率构成，具体见表 1。

Table 1. Composition of financial indicators
表 1. 财务指标构成

指标分类	指标名称	指标分类	指标名称
基本财务指标	流动资产总计 应付账款 资产总计 普通股资本总计 现金和短期投资 销货成本 流通的普通股资本总计 流动负债总计 长期债务发行 长期债务总计 折旧和摊销 特殊项目收入 库存总计 投资和垫款 短期投资总计 流动负债总计 负债总计 净收入(亏损) 财产、厂房和设备总计 优先股资本总计 留存收益 应收账款总计 销售额 普通股和优先股的销售额 应付所得税 利息和相关费用总计 收盘价	财务比率指标	WC 应计 RSST 应计 应收账款变化 库存变化 软资产占比 折旧指数 现金销售变化 现金保证金变化 资产回报率变化 自由现金流变化 留存收益占总资产的比例 总资产的息税前利润 实际发行量 进入市场的订单量

5.2. 识别模型的训练

依据 Bao [13]对数据集的拆分，考虑 2008 年后美国政策和经济形势发生变化，本文选取 1990~2002 年的数据作为模型训练样本集，2003~2008 年的数据作为测试集，用训练样本集训练财务造假识别模型。数据训练环境为 kaggle 云平台，python3.6 环境及一些公开源包，如 sklearn、matplotlib 等。

5.2.1. XGBoost 模型训练

考虑到模型的超参数对训练效果的影响，在模型参与大规模训练前，本文选用 sklearn 包中的组件 GridSearchCV 进行参数选择。GridSearcgCV 是调参利器，只需给定参数范围和小样本数据集，就可对模型在小规模数据集上执行 K 折验证和参数的网格搜索，找到优质的参数。针对 XGBoost，本文随机选取训练样本集中 4000 条数据作为 GridSearchCV 的调参样本，执行 n = 5 的 K 折交叉验证，对给定范围的参数值打分，最终得到范围内最优参数，见表 2。

Table 2. Optimal hyper parameter of XGBoost
表 2. XGBoost 最优超参数

超参数名称	参数解释	值
max_depth	树的最大深度	5
n_estimators	学习子树的预期数量	100
learning_rate	学习率	0.05
colsample_bytree	划分节点时特征的采样比例	0.8
max_delta_step	树权重增量的最大步长	1

将表 GridSearchCV 搜索的最优参数设定为 XGBoost 模型的训练超参数，然后以 1992~2002 年的上市公司的 42 个财务指标特征作为模型输入变量，造假标签“misstate”数据作为模型输出，进行模型训练。

5.2.2. CatBoost 模型训练

同样，在 CatBoost 模型训练前也需要对超参数进行调优。选取 4000 条数据作为调参样本，执行 n = 5 的 K 折交叉验证和打分，得到 CatBoost 模型的最优超参数，见表 3。确定的超参数作为 CatBoost 模型的训练超参数，然后输入样本进行训练。

Table 3. Optimal hyper parameter of CatBoost
表 3. CatBoost 最优超参数

超参数名称	参数解释	值
n_estimators	学习子树的预期数量	100
learning_rate	学习率	0.05
rsm	划分节点时特征的采样比例	0.5

5.3. 结果分析

经训练后的模型，在测试集中检验财务造假识别能力。本文首先从效率和精度上，对模型训练的结果分析。表 4 中模型训练时间显示了 XGBoost 和 CatBoost 的模型训练耗时，XGboost 用时明显低于 CatBoost，因此在训练效率上 XGboost 模型优于 CatBoost 模型。同时准确度表现模型在测试集中对财务造假企业的识别能力，在所有的测试集中，经训练的 CatBoost 模型能够识别约 76.5%的财务造假企业，而 XGBoost 模型能识别出约 72.6%的财务造假企业，XGBoost 模型的准确率略低于 CatBoost 模型。

Table 4. Efficiency and accuracy of the model
表 4. 模型的效率和精度

模型名称	训练时间/min	准确度
XGBoost	0.8	0.726
Catboost	1.2	0.765

其次，对于财务造假识别这一二分类问题，AUC-ROC 曲线能够很好地描绘模型的性能，ROC 对应的 AUC 值越高，分类模型做出正确判定的可能性越大，犯错误的概率越小。由图 2 可知，XGBoost 模型

AUC 为 0.61，CatBoost 模型 AUC 为 0.63，CatBoost 模型性能略强于 XGBoost。

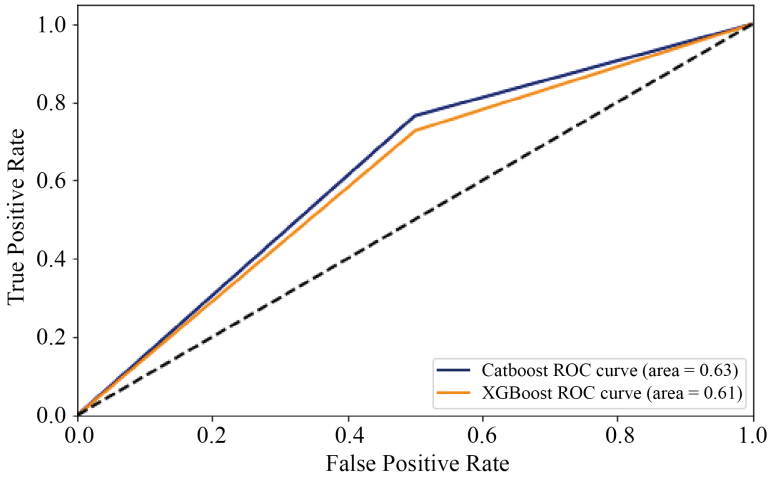

Figure 2. Curve of AUC-ROC
图 2. AUC-ROC 曲线

综合上述分析，CatBoost 模型除了在训练耗时上略长之外，在对财务造假样本的识别方面，准确度和性能上都较优于 XGBoost 模型。

6. 结束语

针对财务造假识别，本文使用 XGBoost 和 CatBoost 这两类 boosting 方法对比分析。以 Yang Bao 等整理公布的 1990~2014 年美国上市公司财务造假数据为基础，划分训练集和测试集，作为两类 boosting 方法模型的数据源。

在对数据源进行大规模训练前，本文首先随机采样少量数据，运用 GridSearchCV 预训练模型，对采样数据进行 K 折交叉验证和网格化参数寻优，找出两类 boosting 模型的最优初始超参数。然后设定 XGBoost 和 CatBoost 模型的初始参数为上述最优超参数，进行大规模训练及测试。结果表明，对于美国上市公司的财务造假识别，两类 boosting 方法均有良好表现，但 CatBoost 方法从识别准确度和性能上略优于 XGBoost 模型。

此外，本文也有一些局限。考虑到政策和经济形势的变化，本文选择了 1990~2008 年的财务造假数据作为数据源，对于 2008 之后的数据，两类 boosting 方法的表现还需进一步研究。在两类模型的超参数选择方面，本文使用随机采样的方式，通过预训练模型进行超参数寻优，有一定的不确定性，仍有一定的优化空间。

参考文献

[1] Altman, E. (1968) Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *Journal of Finance*, **23**, 589. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[2] Messod, B.D. (1999) The Detection of Earnings Manipulation. *Financial Analyst Journal*, **5**, 22-36. https://doi.org/10.2469/faj.v55.n5.2296

[3] Tutino, M. and Merlo, M. (2019) Accounting Fraud: A Literature Review. *Risk Governance and Control*: *Financial Markets and Institutions*, **9**, 8-25. https://doi.org/10.22495/rgcv9i1p1

[4] Qin, R. (2021) Identification of Accounting Fraud Based on Support Vector Machine and Logistic Regression Model. *Complexity*, **2021**, Article ID 5597060. https://doi.org/10.1155/2021/5597060

[5] Zheng, Y., Ye, X. and Wu, T. (2021) Using an Optimized Learning Vector Quantization-(LVQ-) Based Neural Network in Accounting Fraud Recognition. *Computational Intelligence and Neuroscience*, **2021**, Article ID 4113237. https://doi.org/10.1155/2021/4113237

[6] Tu, Z. (2005) Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering. 10*th IEEE International Conference on Computer Vision* (*ICCV*'05), Vol. 1, 1589-1596.

[7] Chen, T. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. P*roceedings of the* 22*nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, New York, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[8] Prokhorenkova, L., Gusev, G., Vorobev, A., *et al*. (2017) CatBoost: Unbiased Boosting with Categorical Features. arXiv preprint arXiv:1706.09516

[9] Paleczek, A., Grochala, D. and Rydosz, A. (2021) Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection. *Sensors*, **21**, 4187. https://doi.org/10.3390/s21124187

[10] Thongsuwan, S., Jaiyen, S., Padcharoen, A., *et al*. (2021) ConvXGB: A New Deep Learning Model for Classification Problems Based on CNN and XGBoost. *Nuclear Engineering and Technology*, **53**, 522-531. https://doi.org/10.1016/j.net.2020.04.008

[11] Shahriar, S.A., Kayes, I., Hasan, K., *et al*. (2021) Potential of Arima-Ann, Arima-Svm, Dt and Catboost for Atmospheric Pm2. 5 Forecasting in Bangladesh. *Atmosphere*, **12**, 100. https://doi.org/10.3390/atmos12010100

[12] Zhang, F,. Yang, J., Liang, B.S., *et al*. (2021,) Analysis of Influencing Factors of New Energy Vehicle Satisfaction Based on Scenario Thinking and Catboost Model. *IOP Conference Series*: *Earth and Environmental Science*, **769**, 042023. https://doi.org/10.1088/1755-1315/769/4/042023

[13] Bao, Y., Ke, B., Li, B., *et al*. (2020) Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach. *Journal of Accounting Research*, **58**, 199-235. https://doi.org/10.1111/1475-679X.12292