集成机器学习模型在不平衡样本财务预警中的应用*

张　露[1]，刘家鹏[1]，江敏祺[2]
（1.中国计量大学 经济与管理学院，浙江 杭州 310018；2.上海财经大学 信息管理与工程学院，上海 200000）

摘　要： 基于上交所主板市场A股企业的财务指标数据来预测企业的财务风险，样本数据包括1 227家正常上市企业和42家被财务预警的企业，数据严重不平衡，通过重采样技术解决了分类器在不平衡样本中失效的问题，运用Bagging思想的集成机器学习对预测模型进行提升与优化。正确挑选出有财务危机企业的概率最高达到92.86%，在此基础上，样本的整体准确率在经过模型的集成之后提高了5.4%。集成模型提高了对上市企业的财务预警能力，能为企业的正常经营和投资者的安全投资提供一定的借鉴。

关键词： 财务预警预测；集成机器学习；不平衡采样技术
中图分类号： TN99；TP391　　**文献标识码：** A　　DOI：10.16157/j.issn.0258-7998.201234

中文引用格式： 张露，刘家鹏，江敏祺. 集成机器学习模型在不平衡样本财务预警中的应用[J].电子技术应用，2021，47(8)：34-38.
英文引用格式： Zhang Lu，Liu Jiapeng，Jiang Minqi. The application of the integrated machine learning model in the financial crisis of imbalanced sample[J]. Application of Electronic Technique，2021，47(8)：34-38.

The application of the integrated machine learning model in the financial crisis of imbalanced sample

Zhang Lu[1]，Liu Jiapeng[1]，Jiang Minqi[2]
（1.School of Economics and Management，China Jiliang University，Hangzhou 310018，China；
2.School of Information Management and Engineering，Shanghai University of Finance and Economics，Shanghai 200000，China）

Abstract： This paper forecast the financial risk of enterprises based on the financial index data of A-share enterprises in the main board market of Shanghai Stock Exchange.The samples included 1227 normal listed enterprises and 42 enterprises which have been financial warning. The data was seriously unbalanced. The problem of classifier failure in unbalanced samples was solved by resampling technology in some certain.The integrated machine learning based on Bagging was used to improve and optimize the prediction model.The highest probability of correctly selecting enterprises with financial warning was 92.86%. On this basis, the overall accuracy of the sample was improved by 5.4% after the integration of the model. The integrated model improved the financial early warning ability of listed enterprises which could provide some reference for the normal operation of enterprises and the safety investment of investors.

Key words： financial early warning prediction；integrated machine learning；imbalanced sampling technology

0 引言

进入大数据时代以来，对信息的敏感程度和预测能力变得尤为重要，而对企业而言，无论是在经营活动还是投资活动中，财务危机预警一直是个问题和难题。机器学习的兴起为大数据的处理和应用提供了新的方式。

目前，许多学者将机器学习与金融危机预警相结合，取得了重大突破。OHLSON J A[1]建议将逻辑回归应用于分类的后概率，来估计公司的破产概率。Zou Hui和HASTIE T[2]提出了弹性网络，克服了岭回归和Lasso的缺点[3]。决策树学习是一种强大的分类器[4]，在树分类器的基础上，有学者提出了随机森林[5]和XGBoost[6]，在计算机[7]、图像分类[8]等领域被证明有效。

但在过去的研究中，大多采用人工设定样本量，而忽视了实际上财务预警企业与正常企业的数量对比的悬殊[9]。数据不平衡的问题是财务预警研究领域的难题[10]。VEGANZONES D和SEVERIN E[11]提出采样技术可用于提高不平衡样本预测的分类器性能，随机上采样技术[12]、随机下采样技术[13]和人工合成少数抽样技术(SMOTE)[14]的应用解决了集成复杂分类器在不平衡的财务预警研究数据中失效的问题。而集成学习机制可以通过集成不同的模型来整合多种算法的优点[15]，目前在个人信贷领域已经有了一定的应用[16]。

* 基金项目：国家自然科学基金(18BGL224)

本文研究的目的包括三个部分：一是测试集成机器学习模型的预测性能，寻找最适合财务预警的分类器；二是将不平衡学习理念运用到中国上市公司的全样本中，避免人工筛选样本的巧合性，利用抽样技术和袋装(Bagging)方法提高企业在T-3期间内财务风险的概率；三是保持财务预警企业预测准确率的同时，提高健康企业分类的准确性，为企业的日常经营和投资者的投资决策提供一定的参考。

1 实证研究方法设计

1.1 研究模型设计

本文的研究模型设计过程如图1所示。

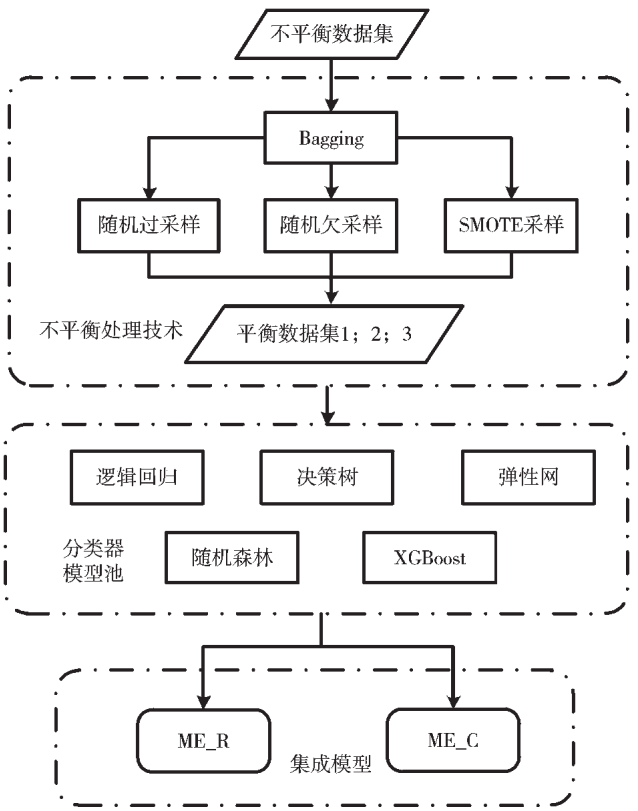


图1 不平衡财务预警模型设计过程

首先，本文用装袋法和采样技术对不平衡数据进行处理。随机上采样技术(RUT)通过随机抽取重复的小样本来平衡不平衡样本；随机下采样技术(RDT)随机筛选出大样本，使其处于平衡状态；合成少数过采样技术(SMOTE)通过KNN生成新的小样本来生成平衡数据，分别得到3个数据集。

其次，对于在上一步骤得到的数据集，分别采用模型池中的Logistic回归(LR)、弹性网(EN)、决策树(DT)、随机森林(RF)和XGBoost 5种分类器进行预测。前4种财务方法在财务预警领域已经有了较为成熟的应用。XGBoost于2016年提出，是对GBDT的进一步提升，其损失函数为：

$$\mathrm{Obj}(\theta)=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{1}$$

其中，第一部分表示n个样本的损失函数值，在这里通过样本预测值$\hat{y}_i$和真实值y_i的比较，来计算出对样本i的模型损失值；第二部分是正则项，用来控制模型的复杂度，模型越复杂，则惩罚力度越大，从而提升模型的泛化能力，$\Omega(f_k)$代表第k棵树的复杂度。XGBoost是一种改进的GBDT算法，GBDT在优化时只用到一阶导数，而XGBoost则对损失函数进行了二阶泰勒展开，利用二阶导加快了模型训练时的收敛速度，使得模型求解更加高效。XGBoost算法中加入了正则项，可以有效减少过拟合，即：

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{2}$$

其中，T为叶节点的个数。第二部分为节点权重的L2范式，叶子节点值w_j用来评估第k棵树的复杂性程度。γ、λ分别为对应的惩罚参数，越大的γ和λ对应越简单的模型。对式(1)泰勒展开，可得：

$$\begin{aligned}\mathrm{Obj}^{t}&\approx\sum_{i=1}^{n}\left[L(y_i,\hat{y}_i)+g_if_i(x_i)+\frac{1}{2}h_if(x_i)\right]+\sum_{k=1}^{K}\Omega(f_i)\\&=\sum_{i=1}^{n}\left[g_iw_{q(x_i)}+\frac{1}{2}h_iw_{q(x_i)}\right]+\gamma T+\frac{1}{2}\sum_{i=1}^{n}w_j^2\\&=\sum_{j=1}^{T}\left[\sum_{i\in I_j}g_i+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma T\\&=\sum_{j=1}^{T}\left[G_jw_j+\frac{1}{2}(H_i+\lambda)w_j^2\right]+\gamma T\end{aligned} \tag{3}$$

式(3)中涉及的参数有：

$$g_i=\partial_{\hat{y}^{t-1}}L(y_i,\hat{y}^{t-1}) \tag{4}$$

$$h_i=\partial^2_{\hat{y}^{t-1}}L(y_i,\hat{y}^{t-1}) \tag{5}$$

$$I_j=\{i|q(x_i)=j\} \tag{6}$$

$$G_j=\sum_{i\in I_j}g_i \tag{7}$$

$$H_j=\sum_{i\in I_j}h_i \tag{8}$$

其中，h_i和g_i为第t步的损失函数，由于h_i和g_i可以并行计算，极大地提高了XGBoost的建模效率；I代表了每个叶子节点上的训练集样本。此外，XGBoost算法还在目标函数中加入了正则项，用以权衡目标函数的下降和模型的复杂程度，一定程度上避免了过拟合。

最后，集成学习机制通过整合不同的学习模型，综合多种算法的优点。本文分别通过稳健和谨慎的算法来整合单个分类器。稳健集成算法是指只要其中一个模型预测到企业的财务风险，集成模型就预测出企业存在财务风险，并记为ME-R；谨慎集成算法是只有所有模型都预测到企业的财务风险时，该集成模型才能预测到企业存在财务风险两个分类器同时预测企业将面临风险，记为ME-C。

1.2 数据来源及指标选取

本文选取的是上交所主板市场非金融行业A股企

业的财务指标数据，数据来自锐思金融数据库。考虑到ST或*ST的标志是连续两年或三年净利润为负，因此选取了$t-3$年的财务指标数据来预测第t年的结果。

本文从锐思金融数据库的财务比率数据中选取了107个原始变量，并参考了数据库的分类方法，将107个变量分成了9组指标，分别是每股指标、盈利能力、偿债能力、成长能力、营运能力、现金流量、分红能力、资本结构和杜邦分析指标。由于这些指标未经过初始分类，存在一定的相关性，为了防止信息冗余和过度拟合，本文采用相关系数矩阵计算，筛选掉相关系数大于0.5的指标，然后剩下的57个变量指标如图2所示，X_i代表财务预警指标。

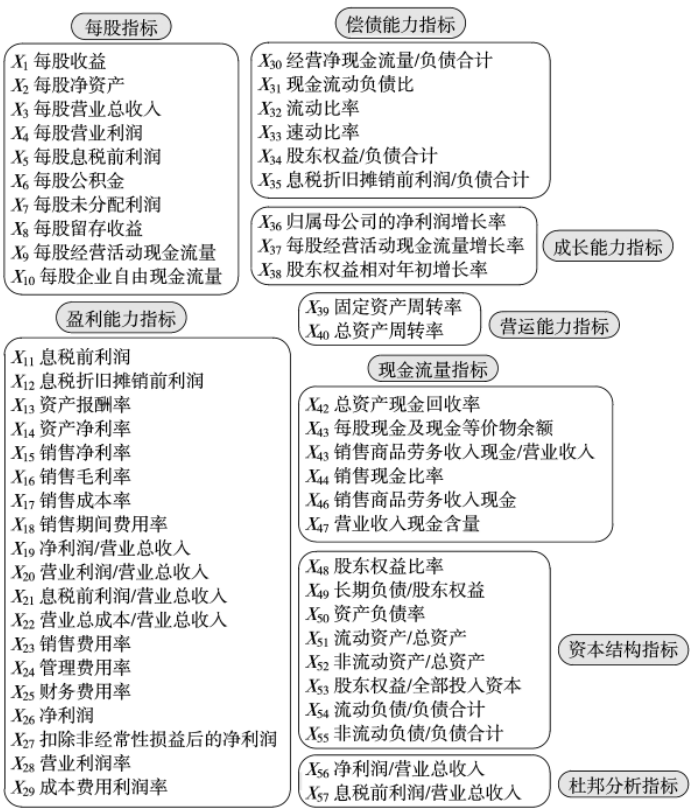


图2 财务预警指标构建

1.3 数据来源及指标选取

本文设定的分类结果矩阵表示如表1所示，TP和TN代表预测和真实值一致的情况，FP和FN代表预测值和真实值不一致的情况。本次研究中感兴趣的是发生财务预警的企业，因此将其设定为Positive的类别。

表1 分类矩阵

	预测值为Positive	预测值为Positive
预测值为Positive	TP	FN
预测值为Negative	FP	TN

表1中，TN代表正确的分类为不感兴趣的类别，TP代表正确的分类为感兴趣的类别，FN代表错误的分类为不感兴趣的类别，FP代表错误的分类为感兴趣的类别。本文使用的3个指标公式如下所示：

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$\text{Specifity}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{10}$$

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{11}$$

其中，灵敏度(Sensitivity)是本文感兴趣的类别正确分类的概率，即正确挑选出有财务风险企业的概率；特异性(Specificity)度量了挑选出正常企业的概率；准确度(Accuracy)则是所有企业被正确分类的概率。

此外，还将用AUC(Area Under Curve)值来度量模型的精确度以衡量模型的性能。AUC值越大，代表该模型的性能越好。

2 实际测试及结果分析

首先使用Bagging的思想加强学习感兴趣样本的信息，然后在Bagging的基础上，又分别尝试使用了随机过采样、随机欠采样和SMOTE采样技术。对上述优化是否能提升模型性能用AUC值来表示，如表2所示，即模型经过优化前后的AUC值的对比。

表2 模型优化前后的AUC值

采样技术	LR	EN	DT	RF	XGBoost
均衡样本	0.542 1	0.758 2	0.707 0	0.772 9	0.809 5
随机欠采样	0.760 1	0.824 6	0.837 6	0.954 1	0.881 1
随机过采样	0.750 8	0.849 7	0.787 7	0.969 6	0.952 9
SMOTE采样	0.785 9	0.808 6	0.845 0	0.962 0	0.944 6

从表2中可以看出，经过采样技术和Bagging对机器学习模型的优化，AUC值得到了明显的提高，分类器在优化前的均衡样本中的表现要明显差于优化后的不均衡样本。数据的增加使得分类器能学习到更多的信息，对样本进行不平衡采样的处理，使得模型不会忽略小样本中的信息，甚至通过权重影响，更重视小样本中的信息，从而减小巧合，发挥分类器预测的性能。

接下来分别对经过不平衡采样处理后的分类器进行财务预警预测，结果如表3所示。

基于误判的代价，本文优先考虑模型的灵敏度，即正确挑选出财务预警企业的概率。其中，在Bagging RDT的算法下，对财务预警的预测准确率是最高的，且随机森林和XGBoost的Sensitivity值是相同的。对此，推测将这两个分类器进一步集成可能会提高整体样本的准确率。因此，本文尝试用稳健和谨慎的算法将随机森林和XGBoost相结合。

从表3中的ME-R和ME-C可以看出，两种集成算法都能保持金融危机企业选择的准确性，但谨慎的集成算法可以降低对健康企业的误判率。在Bagging RDT模型上，总精度提高了5%~9%。因此，推荐谨慎算法(ME-C)

表 3 优化模型的分类预测概率

模型方法	Only Bagging			Bagging RDT			Bagging RUT			Bagging SMOTE		
	灵敏度	特异性	准确度	灵敏度	特异性	准确度	灵敏度	特异性	准确度	灵敏度	特异性	准确度
LR	0.250 0	0.953 4	0.914 0	0.857 1	0.709 7	0.718 0	0.571 4	0.930 1	0.910 0	0.785 7	0.786 0	0.786 0
EN	0.214 3	0.987 3	0.944 0	0.750 0	0.790 3	0.788 0	0.714 3	0.853 8	0.846 0	0.714 3	0.824 2	0.818 0
DT	0.250 0	0.976 7	0.936 0	0.892 9	0.724 6	0.734 0	0.714 3	0.889 8	0.880 0	0.750 0	0.845 3	0.840 0
RF	0.785 7	0.987 3	0.976 0	0.928 6	0.783 9	0.792 0	0.785 7	0.991 5	0.980 0	0.857 1	0.879 2	0.878 0
XGBoost	0.785 7	0.976 7	0.966 0	0.928 6	0.819 9	0.826 0	0.785 7	0.972 5	0.962 0	0.857 1	0.906 8	0.904 0
ME-R	0.785 7	0.970 3	0.944 0	0.928 6	0.728 8	0.740 0	0.785 7	0.972 5	0.962 0	0.857 1	0.862 3	0.862 0
ME-C	0.785 7	0.991 5	0.980 0	0.928 6	0.877 1	0.880 0	0.785 7	0.991 5	0.980 0	0.857 1	0.923 7	0.920 0

下的集成模型。

此外，通过随机森林和 XGBoost 对研究指标进行重要性分析，分别排名前 5 个的变量如图 3 所示挑选出重要指标，为利益相关者提供一定的参考，如图 3 所示。

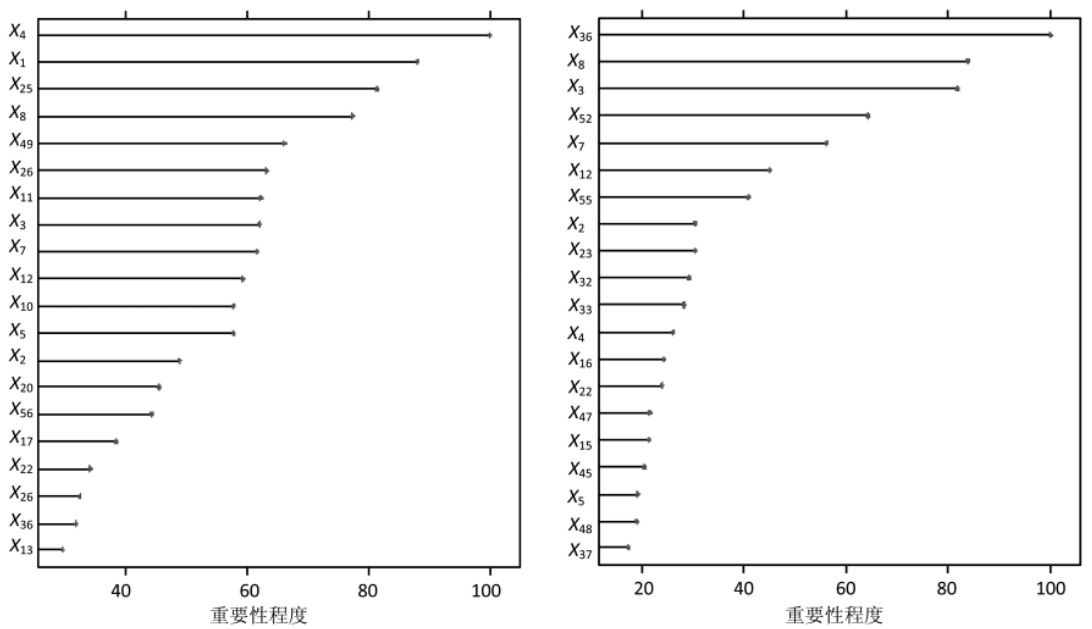
在图 3 中有一个变量发生重叠，因此，一共有 9 个较为重要的变量，分别是每股收益、每股营业总收入、每股营业利润、每股未分配利润、每股留存收益、归属母公司的净利润增长率、每股现金及现金等价物余额、流动负债/负债合计、扣除非经常性损益后的净利润。筛选出的衡量企业财务风险的关键性指标，能为企业的投资决策和经营管理提供一定的借鉴。

3 结论

本文将集成机器学习模型应用到不均衡样本的企业财务预警中，并通过一系列的优化解决了样本不均衡的问题，提高了预测的准确性。

本文的实证研究使用了 $t-3$ 期的上交所主板市场非金融行业 A 股企业的财务指标数据来预测 t 期的企业财务状况，即预测该企业在 t 期是否会被 ST。本文证明了不同的采样比例会影响预测的准确率，随着样本规模的增大，在一定程度上会提高预测准确率，但随着正常上市企业样本的扩增，而存在财务风险的企业的数量远远小于正常上市企业，使得分类器“偷懒”，盲目将企业预测为正常，出现了样本的不均衡现象，使得模型失去挑选出财务危机企业的能力。但是由于人为设定样本使得样本量数量受限，使得机器学习的分类器无法完全发挥其优势，因此本文应用了 Bagging 思想和采样技术——随机过采样、随机欠采样和 SMOTE 采样来优化模型，从而提升预测的准确性。

实证研究表明，采样技术的使用提高了模型的性能，提升了正确挑选出财务预警企业的概率，这正是本文所感兴趣的分类。其中，单独的分类器中，表现最佳的是 XGBoost 与随机欠采样的结合，它在提升了挑选出财务危机企业的概率的同时，对正常企业预测概率的兼顾性要优于随机森林。为了减少正常企业被误判的概率，本文对随机森林和 XGBoost 进行了简单的集成，使得在 t 期正确预测财务预警企业的概率维持在 92.86% 的同



(a)随机森林　　(b)XGBoost

图 3 随机森林的指标重要性程度

时，相比于基分类器，集成模型将正常企业的误判率降低了约6%，整体预测准确率提高了5.4%。

集成机器学习的应用能帮助企业较好地完成前瞻性的财务预警，与传统方法相比，会具有更好的普适性，能结合大数据时代的背景，提高预测的准确率，对管理者有更低的财会专业性要求，有利于企业的多元化发展，为企业挑选投资对象以及日常的生产经营活动提供了新的借鉴意义。

参考文献

[1] OHLSON J A.Financial ratios and the probabilistic prediction of bankruptcy[J].Journal of Accounting Research, 1980(18): 109-131.

[2] Zou Hui, HASTIE T.Regularization and variable selection via the elastic net[J].Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2005, 67(2): 301-320.

[3] 宋瑞琪，朱永忠，王新军.高维数据中变量选择研究[J].统计与决策，2019，35(2)：13-16.

[4] 何元.基于云计算的海量数据挖掘分类算法研究[D].成都：电子科技大学，2011.

[5] BREIMAN L.Random forests[J].Machine Learning, 2001, 45(1): 5-32.

[6] CHEN T, GUESTRIN C.Xgboost: a scalable tree Boosting system[C]//Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining.ACM, 2016: 785-794.

[7] 高洪波，李磊，周婉婷，等.基于XGBoost的硬件木马检测方法[J].电子技术应用，2019，45(4)：55-59.

[8] 肖玉玲，仵征，朱煜.结合分块模糊熵和随机森林的图像分类方法[J].电子技术应用，2017，43(7)：122-126.

[9] 李清，于萍.财务危机预测主要方法比较研究[J].数理统计与管理，2012，31(4)：689-706.

[10] 李扬，李竟翔，马双鸽.不平衡数据的企业财务预警模型研究[J].数理统计与管理，2016，35(5)：893-906.

[11] VEGANZONES D, SEVERIN E.An investigation of bankruptcy prediction in imbalanced datasets[J].Decision Support Systems, 2018(8): 111-124.

[12] HE H, GARCIA E A.Learning from imbalanced data[J].IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9): 1263-1284.

[13] TAHIR M A, KITTLER J, MIKOLAJCZYK K, et al.A multiple expert approach to the class imbalance problem using inverse random under sampling[J].Multiple Classifier Systems, Springer, 2009(6): 82-91.

[14] CHAWLA N V, BOWYER K W, HALL L O, et al.SMOTE: synthetic minority over-sampling technique[J].Journal of Artificial Intelligence Research, 2011, 16(1): 321-357.

[15] WANG B, PINEAU J.Online Bagging and Boosting for imbalanced data streams[J].IEEE Transactions on Knowledge & Data Engineering, 2016, 28(12): 3353-3366.

[16] Wang Chongren, Han Dongmei.Personal credit evaluation of Internet credit based on hyperparameter optimization and integrated learning[J].Statistics and Decision-Making, 2019, 35(1): 89-93.

（收稿日期：2020-12-23）

作者简介：

张露(1995-)，女，硕士研究生，主要研究方向：机器学习与公司金融。

刘家鹏(1969-)，通信作者，男，博士，教授，主要研究方向：金融信息技术，E-mail：jpliu@126.com。

江敏祺(1995-)，男，博士研究生，主要研究方向：量化投资。

扫码下载电子文档

（上接第33页）

digital power supply noise analyzer with enhanced spectrum measurements[J].IEEE J Sol Sta Circ, 2015, 50(7): 1711-1721.

[6] ALON E, STOJANOVIC V, HOROWITZ M.Circuits and techniques for high-resolution measurement of on-chip power supply noise[J].IEEE J Sol Sta Circ, 2005, 40(4): 820-828.

[7] ALON E, ABRAMZON V, NEZAMFAR B, et al.On-die power supply noise measurement techniques[J].IEEE Trans AP, 2009, 32(2): 248-259.

[8] ZHAI P, ZHOU X, CAI Y, et al.A scalable 20 GHz on-die power supply noise analyzer with compressed sensing[C]//IEEE ISSCC.San Francisco, CA, USA.2020: 386-388.

[9] ZHAI P, ZHOU X, CAI Y, et al.A multi-slice VCO-based quantizer for on-chip power supply noise analysis achieving 0.11(mV)2/sqrt(MHz) noise floor[C]//IEEE ASSCC, Macau, China, 2019: 121-122.

[10] NOURANI M, RADHKRISHNAN A.Power-supply noise in SoCs: ATPG estimation and control[C]//IEEE Int conf Test, Austin, TX.2005: 507-516.

（收稿日期：2020-12-17）

作者简介：

翟鹏飞(1991-)，男，博士研究生，主要研究方向：电源完整性和模拟集成电路设计。

周雄(1987-)，男，博士，副教授，主要研究方向：模拟及数模混合信号集成电路设计。

李强(1979-)，男，博士，教授，主要研究方向：模拟及数模混合信号集成电路设计。

扫码下载电子文档

版权声明