Statistics and Application 统计学与应用, 2022, 11(1), 150-156
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111017

Hans汉斯

基于改进随机森林算法的上市公司信用风险实证分析

王晓筱，王永茂

燕山大学理学院，河北 秦皇岛

收稿日期：2022年1月17日；录用日期：2022年2月4日；发布日期：2022年2月17日

摘　要

近年来随着金融市场的不断发展，贷前识别风险企业、有效进行信贷风险控制越来越重要。本文主要研究企业信用风险评估的问题，通过合理的模型选择及模型优化，提升模型识别问题企业的能力。本文首先基于实际情况选择了合理的模型评估指标体系，通过优化后的随机森林算法，将特征选取与模型训练过程相结合，利用该模型以我国上市公司数据为例，进行了实证检验，并横向对比常见评估模型的数据表现，实验结果表明模型有较好的预测效果。

关键词

信用风险评估，随机森林，特征递归消除法

Empirical Analysis of Credit Risk of Listed Companies Based on Improved Random Forest Algorithm

Xiaoxiao Wang, Yongmao Wang

School of Science, Yanshan University, Qinhuangdao Hebei

Received: Jan. 17th, 2022; accepted: Feb. 4th, 2022; published: Feb. 17th, 2022

Abstract

With the continuous development of the financial market in recent years, it has become more and more important to identify risky enterprises before lending and effectively control credit risks. This paper mainly studies the problem of enterprise credit risk assessment, and improves the ability

文章引用: 王晓筱, 王永茂. 基于改进随机森林算法的上市公司信用风险实证分析[J]. 统计学与应用, 2022, 11(1): 150-156. DOI: 10.12677/sa.2022.111017

of the model to identify problem enterprises through reasonable model selection and model optimization. This paper first selects a reasonable model evaluation index system based on the actual situation, and combines the feature selection and model training process through the optimized random forest algorithm. Comparing the data performance of common evaluation models, the experimental results show that the model has better prediction effect.

Keywords

Credit Risk Assessment, Random Forest, Feature Recursive Elimination

1. 引言

进入 21 世纪以来，随着经济社会的飞速发展，金融市场的健康稳定对于促进国家经济发展的意义越来越重要。在此背景下，怎样选择合适的评估模型和指标系统、科学地衡量企业信用风险，具有高度的理论意义和现实意义。

企业的信用风险评估体系的建立主要分成两部分，首先是建立合适的指标体系，通过指标可以明确企业的信用情况，以及进行借贷的风险性大小；其次是结合信用风险指标体系建立合适的评估模型，通过模型可以在贷前能够识别风险企业，协助金融机构建立行之有效的风险管理机制[1]。

近年来，随着大数据时代的到来，相较于传统评估方式强依赖于专家意见，现阶段评估方法多为参考专家意见与业务场景建立评估模型。1993 年张更生、蒯本江提出建立信贷风险预警体系，系统阐释了信贷风险的来源，以及信贷风险的九类特征，包括资金流动比率、负债比率资产流失比率等[2]。张雷，王家琪等人提出基于 RF-SMOTE-XGboost 的银行用户个人信用风险评估模型，所建立的模型在评估时具有更好的精度与收敛性[3]。

时至今日，信用风险评估领域的研究已十分丰富，但此类问题涉及情况较为复杂、特征繁多，在模型的训练过程中，仍然存在着过拟合，特征选取困难等诸多问题。本文选择优化后的随机森林算法，利用随机森林算法降低过拟合风险，将特征选取与模型训练过程相结合，建立了基于随机森林算法的信用评估模型，并以某金融信贷机构披露的我国上市公司数据为例，测试了模型的准确性，一定程度上解决了上述问题，有效提高了预测精度。

2. 理论基础

2.1. 随机森林

随机森林最早是由 Breiman 等多位学者共同提出的一种机器学习算法。原理可看作从原始训练样本集中有放回地重复随机抽取 K 个样本生成新的训练样本集合，训练多个决策树 $\{h(X,\theta_k),k=1,2,3,\cdots,K\}$ 共同参与分类决策的组合模型，$\{\theta_k\}$ 为服从独立同分布的随机变量，完成训练后得到分类模型序列 $\{h_1(X),h_2(X),h_3(X),\cdots,h_k(X)\}$，再用它们构成一个多分类模型系统，最终分类结果采用简单多数投票法[4]。最终的分类决策：$H(x)=arg\max_y\sum_k I(h_k(x)=Y)$，其中 $H(x)$ 表示组合分类模型，h_i 是单个决策树分类模型，Y 表示输出变量(或称目标变量)，$I(\cdot)$ 为示性函数。

由于随机森林在训练的时候，每一棵树的输入样本都不是全部的样本，对于特征的选择也同样是随机的，两次随机采样的过程使得随机森林很大程度上避免了过拟合的问题。除此之外，两个随机性的引入也使得模型具有较好的抗噪能力。

2.2. 基于特征递归消除的随机森林模型

随递归特征消除(RFE)是 Guyon 等人基于支持向量机提出的。主要原理是将特征集合初始化为整个数据集合，每次剔除一个排序准则分数最小的特征，直到获得最后的特征集来达到特征筛选的目的[5]。本文选择了基于特征递归消除法的随机森林模型即改进后的随机森林模型，将随机森林模型与递归特征消除法相结合，将指标的筛选与模型的训练融合，具体步骤如下：

假设第 i 个特征的排序准则分数定义为：

$$c_i = w_i^2$$

每次迭代中去除排序准则分数最小的特征，然后运用剩余的特征训练随机森林，进行下一次的迭代。具体算法步骤如下：

步骤 1：原始数据进行预处理，把集合 X 变为

$$X_j = \begin{cases} \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N_1+N_{j+1}}, j = 1,2,\cdots,n-1 \\ \text{当} v_i = 1\text{时}，\ y_i = 1; \\ \text{当} v_i = j+1\text{时}，\ y_i = -1 \\ \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N_2+N_{j-n+3}}, j = n,2,\cdots,2n-3 \\ \text{当} v_i = 2\text{时}，\ y_i = 1; \\ \text{当} v_i = j-n+3\text{时}，\ y_i = -1 \\ \vdots \\ \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N_{n-1}+N_n}, j = n\left(n-1\right)/2 \\ \text{当} v_i = n-1\text{时}，\ y_i = 1; \\ \text{当} v_i = n\text{时}，\ y_i = -1 \end{cases}$$

步骤 2：对数据集合 X_j 构建模型，根据模型准确度进行特征选择，所得到的对应特征引子集为 $F_j \subseteq \{1,2,\cdots,D\}, j = 1,2,\cdots,\left(n\left(n-1\right)\right)/2$；

步骤 3：将得到的特征子集进行 F_j 合并，得到最终的特征子集为 $F = \bigcup_{j=1}^{\left(n(n-1)\right)/2} F_j$，输出由该特征子集所构成的模型。

3. 实证分析

3.1. 数据描述与预处理

实验数据为某金融机构对于 2010~2019 年间 a 股上市公司信用评级以及各公司各类财务数据汇总得到，包含信用评分，资产负债率、企业规模、销售收入增长率等共计 58 个指标，共 34540 条数据，表 1 为部分特征格式枚举。其中信用评分 60 分及以上定义为信用良好，用标签 1 代替；反之则用标签 0 代替。实验所用编程语言为 R 语言。

由于原始数据部分特征存在缺失和异常，为保证模型训练的准确性需要对数据进行预处理。预处理方式主要包括以下三类：

Table 1. Enterprise data format
表 1. 企业数据格式

特征	含义	数据类型	示例
股票代码	资产对数	Int	000001
企业规模	资产对数	Float	27.860
二职合一	企业 CEO 是否兼任董事会主席	Object	Y

去除唯一属性：对于无意义的唯一属性特征进行删减，如股票代码；

处理缺失值：对于大量样本都存在缺失的特征进行删减，对极小部分存在大量特征值缺失的样本进行删减，对少量缺失部分数值型特征的样本进行多重插补；

数据标准化：对各个特征进行归一化处理，统一将数字映射到[0,1]上，处理公式如下

$$v_s = \frac{v - v_{\min}}{v_{\max} - v_{\min}}$$

预处理后共有 2025 条有效数据，剩余 34 个特征，其中正向样本 1746 条，负向样本 279 条。

3.2. 实现过程

数据进行预处理后，基于改进随机森林算法，同时进行模型特征的选取以及模型的训练。将所有样本按照 7:3 的比例进行随机抽样，分别构成训练集与测试集，并利用训练集对模型进行训练，利用测试集对模型的效果进行检验，最后通过合理的模型评价指标对模型的预测效果进行评价，主要流程见图 1。

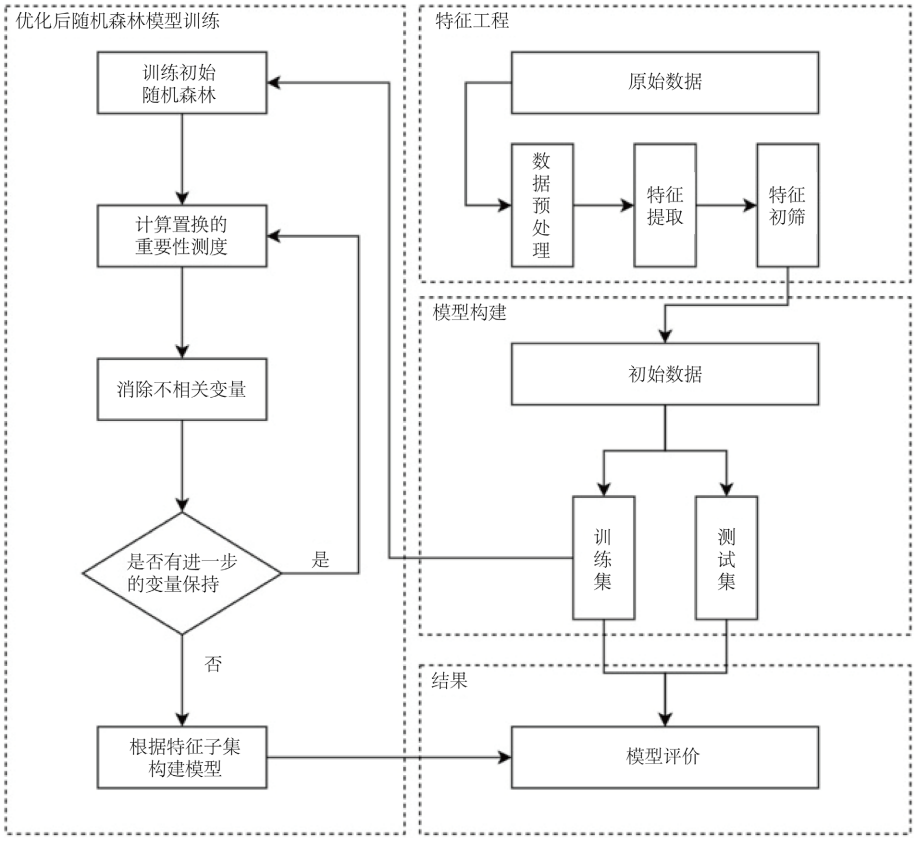

Figure 1. Algorithm flowchart
图 1. 算法流程图

3.3. 实验结果

3.3.1. 评价指标

经过实验数据的预处理，处理后的数据中正负项样本数量存在不平衡的情况，即信用良好的企业样本远多于信用较差企业样本，所以在评价指标的选取上，需要同时关注对于正、负项的分类能力，在样本不平衡的情况下，依然能够对模型做出合理的评价。而 AUC 对样本类别是否均衡并不敏感。

故在众多评价指标中选择 AUC (Area Under Curve)作为核心评价指标，同时选取准确率 A (accuracy)、精准率 P (precision)与召回率 R (recall)作为辅助评价指标。

本文将正常样本作为正样本，违约样本作为负样本，则混淆矩阵定义如下表 2：

Table 2. Confusion matrix definition table
表 2. 混淆矩阵定义表

混淆矩阵		真实	
		正类	负类
预测	正类	TP	FN
	负类	FP	TN

基于混淆矩阵定义的 TP、FN、FP、TN，评价指标定义如下：

AUC 为 ROC 曲线下面积，ROC 曲线核心关注两个指标

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

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

其他指标

$$A = \frac{(TN + TP)}{(FN + FP + TN + TP)}$$

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

$$R = \frac{TP}{(FN + TP)}$$

根据以上指标可以系统评价模型准确性。

3.3.2. 结果分析

再经过模型的训练与测试后，共选取 15 个指标，模型 AUC 为 0.687。进行指标筛选及模型训练的过程中，随着指标数增多 accuracy 变化如图 2 所示，在训练集中，17 个特征入模时，模型 accuracy 达到最高为 0.87。模型 ROC 曲线如图 3 所示。

为对比模型效果，故使用常见的几种分类算法分别建立模型，模型效果如表 3 所示。

如表 3 所示，对比其他单分类器与多分类器模型表现，本文模型 AUC 最高，对比其他模型，有效避免了过拟合的情况，同时提升了对于坏样本的识别能力。Xgboost 模型由于过拟合，出现了 AUC 过低的

情况，而 lda 模型则出现了对于坏样本识别率极低的情况，故证明选取的模型评价指标较为合理，且本文模型有较好的表现。

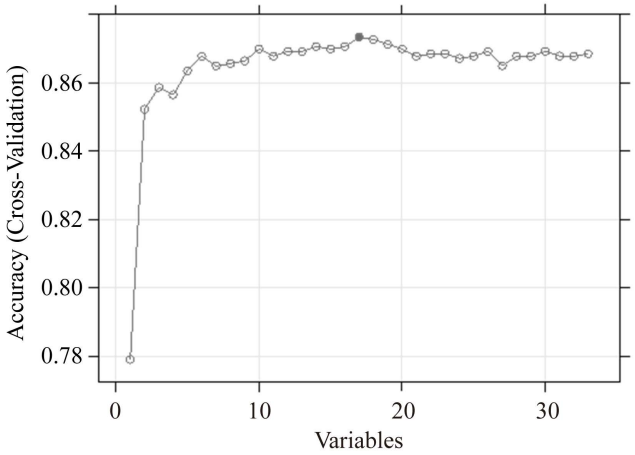

Figure 2. Feature selection based on RFE
图 2. 递归特征消除法特征选择

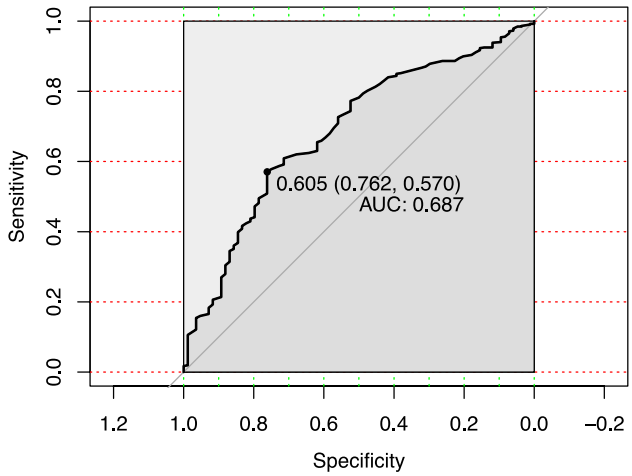

Figure 3. ROC curve
图 3. ROC 曲线

Table 3. Model performance evaluation table
表 3. 模型性能评价表

模型	accuracy	precision	recall	Auc
RFE-RF	0.83	0.97	0.95	0.687
Xgboost	0.76	0.87	0.84	0.535
lda	0.85	0.86	0.98	0.514

4. 结束语

本文选择合理的模型评价指标后，基于改进随机森林的信用风险评估模型，以上市公司数据为例训练模型，并通过对比分析其他常见分类器算法模型表现，验证了该模型对于指标较多的信用风险评估类

问题具有更强的适用性，一定程度上解决了此类模型通常存在的过拟合，以及指标的选取困难的问题，具有一定的理论意义和现实意义。

参考文献

[1] 毛子林, 刘姜. 基于机器学习方法的信用风险评估综述[J]. 经济研究导刊, 2021(23): 117-119.

[2] 张更生, 蒯本江, 韦月斌. 试论建立信贷风险预警体系[J]. 财经理论与实践, 1993(3): 45-47.

[3] 张雷, 王家琪, 费职友, 罗帅, 隋京岐. 基于 RF-SMOTE-XGboost 下的银行用户个人信用风险评估模型[J]. 现代电子技术, 2020, 43(16): 76-81.

[4] 周永圣, 崔佳丽, 周琳云, 孙红霞, 刘淑芹. 基于改进的随机森林模型的个人信用风险评估研究[J]. 征信, 2020, 38(1): 28-32.

[5] 吴辰文, 梁靖涵, 王伟, 李长生. 基于递归特征消除方法的随机森林算法[J]. 统计与决策, 2017(21): 60-63.