计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):222−226 [doi: 10.15888/j.cnki.csa.007866]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 CLPSO-CatBoost 的贷款风险预测方法①

张　涛, 范　博

(北京工业大学 信息学部, 北京 100124)
通讯作者: 范　博, E-mail: fanbo9611@163.com

摘　要: 贷款风险分析是全球金融机构面临的共同考验. 在大数据背景下, 通过机器学习算法预防贷款风险具有现实意义. 针对贷款数据不平衡、噪声大等特点, 本文采用 Boruta 特征选择算法对贷款数据进行重要性筛选;提出通过综合学习粒子群算法 (Comprehensive Learning Particle Swarm Optimization, CLPSO) 优化 CatBoost 集成学习算法 (CLPSO-CatBoost) 的贷款风险预测方法, 该算法改善了全局搜索能力、避免了陷入容易陷入局部最优的问题. CLPSO-CatBoost 相较于传统信用评估模型具有更好的准确性, 有实际应用价值.

关键词: 特征选择; 贷款风险; 综合学习粒子群; CatBoost

引用格式: 张涛,范博.基于 CLPSO-CatBoost 的贷款风险预测方法.计算机系统应用,2021,30(4):222–226. http://www.c-s-a.org.cn/1003-3254/7866.html

Loan Risk Prediction Method Based on CLPSO-CatBoost

ZHANG Tao, FAN Bo

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Loan risk analysis is a common test faced by global financial institutions. In the context of big data, it is of practical significance to prevent loan risks through machine learning algorithms. Aiming at the imbalance in loan data and high noise, this study uses the Boruta feature selection algorithm to sort the importance of loan data. In addition, it proposes the CatBoost integrated learning algorithm based on Comprehensive Learning Particle Swarm Optimization (CLPSO-CatBoost) for loan risk prediction. This algorithm improves the global search and avoids the local optimum. Compared with the traditional credit evaluation models, CLPSO-CatBoost has high accuracy.

Key words: feature selection; loan risk; CLPSO; CatBoost

随着全球金融市场的飞速发展,小微贷款成为金融机构的一项重要业务. 随着贷款空间的极速膨胀, 传统金融机构贷款业务范围正不断受到冲击, 使得互联网金融高速发展. 因此, 需要有效地控制风险手段维持行业健康发展. 大量坏账的出现将会产生类似次贷危机的事件, 对我国经济发展造成影响[1]. 随着数据挖掘技术的逐渐发展, 其在金融领域的应用也更加完善. 机器学习技术为贷款风险预测提供了一个统一、多方面且完整的标准, 进而提高贷款分析的效率与质量.

对于贷款风险预测, 国内外已有较为丰富的研究. 传统方法主要根据相应财务指标、个人征信数据、借款意图等多方面结合, 根据专家意见形成模型. 国内外对于信用评估已有丰富的研究, 主要根据一些财务指标计算结合专家意见形成模型. Fernandes 等人通过验证提出了基于 Logistic 回归的信用评估模型, 成为了用户信用评估的主流方法之一[2]. 随后在国内金融环境里, 梁琪分析国内沪深上市公司失败原因, 指出主成分判别模型在风险监测和信用评估上带来很高的应用价值[3]. 然而传统回归模型准确率偏低, 难以达到预测效果. 而今, 结合机器学习技术建模已成趋势. 基于随机

① 收稿时间: 2020-08-06; 修改时间: 2020-08-28, 2020-09-11; 采用时间: 2020-09-18; csa 在线出版时间: 2021-03-30

森林的组合分类算法[4] 被证明在贷款风险预测算法上有着更高的精度和稳定性. 郭春桃[5] 通过比较 6 种中小企业信用风险预测的方法, 证明集成机器学习模型准确率高于单一模型. 贷款数据中包含年龄、性别等多类型数据, 运用 Gradient Boosting 方法有很好的表现. CatBoost (Category Boosting) 是一种基于梯度的提升树算法, 能够很好的处理类别型特征[6]. 其采用组合类别特征, 可有利于发掘特征之间的联系.

粒子群算法 (Particle Swarm Optimization, PSO) 因其较快的收敛速度和简单的算法结构而广泛应用于各大优化问题中[7]. 种群所包含的粒子为问题的解, 通过目标函数选出每次迭代后粒子的个体最优值及种群最优值. 随后通过粒子速度 v 更新各个粒子的位置. 然而, PSO 算法在解决多峰问题时, 容易陷入局部极值, 最终导致提前收敛. Liang 等人提出了综合学习离子群 (Comprehensive Learning Particle Swarm Optimization, CLPSO) 算法[8].

根据贷款信用数据复杂非线性的特点, 本文采用 CatBoost 模型对贷款风险进行预测. 由于 CatBoost 自身部分超参数可解释性较弱, 超参数将会影响模型的准确率. 本文利用综合学习粒子群 (CLPSO) 算法对 CatBoost 进行优化. CLPSO 算法是一种具有优良全局搜索能力的改进粒子群算法. 在一定程度上避免了粒子群算法的局部收敛问题, 可以有效提高 CatBoost 贷款预测模型的精度.

1 CLPSO 与 CatBoost 算法原理

1.1 综合学习离子群优化算法

传统粒子群算法通过自身最优和全局最优两个值来更新速度与位置, 当全局最优值陷入局部极值时, 所有粒子则容易向其学习陷入局部极值中[9]. CLPSO 算法的速度更新公式并未引入单一粒子向全局最优值学习的部分, 而是将所有粒子的P_{best}作为学习样本. 这样可以提高粒子在不同维度间的信息交换, 从而提高了种群多样性. 其中, 个体最优值即单个粒子从开始至当前时间 k 内找到的最优解P_{best}, 全局最优值即所有粒子在当前时间内找到的最优解P_{global}[10]. 粒子 k 在 i 时刻的速度更新公式如下所示:

$$v_k^d = \omega v_{lk}^d + c_1 r_1 \left[P_{\text{best}f_k(d)}^{\ d} - x_{lk}^d \right] \tag{1}$$

式中, v_k^d为粒子 k 在 d 维的速度, k 为粒子编号 n 为粒子总数. ω为在 (0,1] 的惯性权重, v_{lk}^d为上一时刻粒子 k 在 d 维的速度, c_1为在 (0,2] 的学习因子. r_1为在 (0,1] 均匀分布的随机数, x_{lk}^d为 k 粒子上一时刻在 d 维的位置. $P_{\text{best}f_k(d)}^{\ d}$为粒子 k 在 d 维学习样本$f_k(d)$的最优解. $f_k(d)$为根据学习概率P_k选择出的学习样本. 在进行粒子寻优时, 没有向群体最佳经验学习的例子将会通过产生随机数与P_k比较的方式确定学习对象. 若随机数大于P_k, 则$f_k(d)$为粒子自身的最佳经验, 否则向其他例子学习. P_k表达式如下[10]:

$$p_k = 0.05 + 0.45 \frac{\exp\left(\frac{5(i-1)}{n-1}\right) - 1}{e^5 - 1} \tag{2}$$

根据粒子 k 的速度更新公式, 其位置更新公式如下:

$$x_k^d = x_{lk}^d + v_k^d \tag{3}$$

1.2 CatBoost 算法

在进行特征工程时, CatBoost 算法采用了独有的分类模式. 将所有特征按照贪婪策略进行整合, 计算特征的统计特性与出现频率, 并根据自身所设置的超参数产生特定的衍生字段[11].

传统梯度下降决策树 (GBDT) 模型训练弱学习器时均基于相同的数据集求得模型的精度从而导致梯度估计偏差, 最终导致预测偏差[12]. 弱学习算法如下:

$$h_t = \underset{h \in H}{\operatorname{argmin}} E\left(-g^t(x, y) - h(x)\right)^2 \tag{4}$$

式中, h_t为 t 代弱学习器, $g^t(x,y)$为损失函数的梯度. 面对 GBDT 模型共有的预测偏移问题, CatBoost 算法通过提出 ordered boosting 的方式计算损失函数的梯度, 从而得到无偏梯度估计. 对每个样本x_i, 算法通过不包含样本的训练集训练单独的模型M_i. 针对每个样本所获得的模型M_i, 算法采用求取偏差的方式得到关于样本的梯度估计, 克服了预测偏移.

在进行预测时, CatBoost 采用完全平衡树作为基础预测器. 因完全平衡二叉树的对称结构, 其叶子结点索引可编码为二进制向量, 长度等于树的深度. 算法可将所有特征进行二值化用以进行模型的预测.

2 贷款风险评估模型

2.1 特征工程

本次实验数据选用国内某金融企业在 2016 年至

2019 年间的车辆抵押贷款情况, 总计数据 26 657 条, 包含身份信息、车辆信息、社交数据、填写行为多方面总计 78 个特征, 其中正样本 21 068 条, 存在贷款风险的样本 5589 条.

(1) 缺失值处理

现实数据往往因为多种原因导致数据缺失. 实验根据数据的缺失程度对数据进行分别处理诸如车辆保险金额、申请人工作年限等缺失率大于 30% 的特征, 因对实验模型存在较大影响, 将其删除. 对于缺失率低于 30% 的特征, 将使用差补法对其进行补全. 对于连续特征如审核时间, 采取均值插补法, 以特征的中位数进行补充. 对于类型特征如职业、学历等以 NAN 进行补充. 对于离散特征, 如以主要手段为采用均值插补和众数插补的方式, 根据该特征其余数据预测填补. 去除缺失比例较高的特征后, 实验数据剩余 57 个特征.

(2) 独热编码

数量大于 3 的类型特征通常并不存在比较关系, 但仍以 0、1、2 进行表示, 而实验算法可能将其数值大小进行逻辑判断. 因此将类型特征进行独热编码, 以性别为例, 将原本的单一特征转化为是否男性、是否女性两个特征.

对类型变量进行独热编码, 共生成 48 个子变量, 包括性别、学历、收入途径、配偶状态、申请渠道等.

(3) 特征衍生

使用原始特征进行训练所得模型的泛化能力往往较差[13], 且模型很难挖掘特征之间的联系. 故通过传统信用评估经验, 从交易异常程度、用户还款能力两方面构造衍生变量.

其中交易异常程度包括押品价格与贷款比例、渠道商与申请人是否位于同一市区、借款用途与收入途径是否匹配等. 用户还款能力包括近 360 天内用户逾期次数、借款金额与收入比例等.

(4) 特征选择

实验选用 Boruta[14] 算法对特征进行选择. Boruta 是一种随机森林包装器, 可在不调整参数的情况下对数个特征的重要性进行估计, 筛选出与因变量具有相关性的特征集合. Boruta 对每个特征进行计算, 创建出相对应的阴影特征, 根据阴影特征的随机性判断特征的重要程度.

选用 XGBoost 作为 Boruta 的学习估计器[15], 定义 80 分位数为选择阴影与真实特征的比较阈值, 经过特征工程后得到以下包含 4 维 19 个特征的数据集, 如表 1 所示.

表 1 特征信息表

特征维度	特征
身份信息	性别、年龄、学历、收入途径、配偶状态、逾期次数
贷款信息	贷款利息、申请金额、贷款周期、放款金额、用途
押品信息	车龄、行驶里程、车辆性能、车辆外观、市场价格
衍生特征	渠道信息、审核时长、押品价格与贷款比例

2.2 实现流程

通过特征工程, 实验算法已经可以很好地识别数据进行训练, 但仅采用 CatBoost 模型进行训练在精度表现上仍有一定提升空间. 因此使用 CLPSO 算法优化 CatBoost 模型中的超参数, 包含学习速率、正则子参数及贝叶斯套袋控制强度. 将实验数据按照 4:1 的比例随机分为训练集与测试集, 并进行试验, 实验步骤如下: ① 数据归一化, 得到可进行实验的数据集 D; ② 初始化 CLPSO 算法参数, 设定粒子总数k=20, 惯性参数ω=20, 维度d=3, 学习因子c_1=1.5; ③ 使用 CatBoost 训练模型, 完成后带入测试集, 将交叉熵损失 Logloss 作为适应度函数[16]; ④ 粒子群算法迭代, 寻找当适应度函数f最小时, 所使用的最优参数向量; ⑤ 将最优参数带入 CatBoost 模型中训练. 算法流程如图 1.

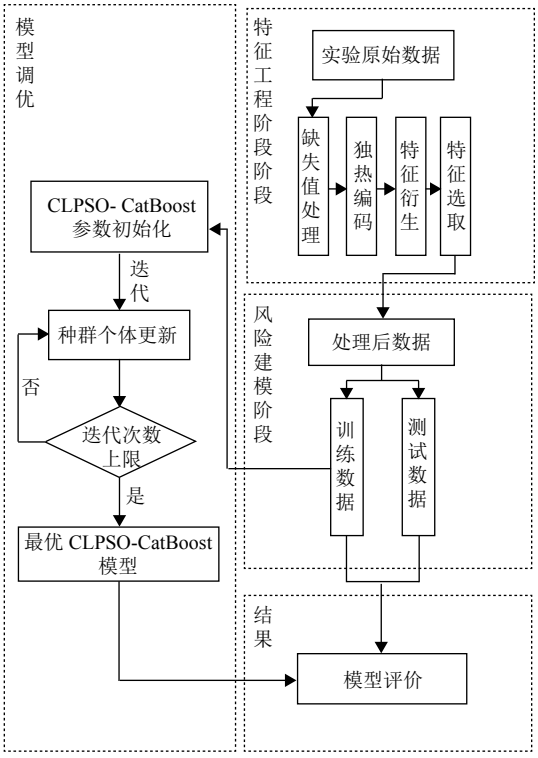

图 1 CLPSO-CatBoost 算法流程图

3 实验分析

3.1 评价指标

贷款数据存在数据不平衡的特性, 即违约风险样本数量远小于正常贷款样本数量. 因此当应用分类准确率作为最终评价指标时, 往往会由于违约样本数量过小导致即使误判违约, 准确率也保持在很高的水平. 这样就失去了贷款违约的预测能力, 然而个别贷款违约在日常生活中往往会对金融公司产生很大的代价. 因此将准确率作为评价指标在预测贷款风险的二分类模型中是不合理的. 本文采用 AUC 值作为评价指标. 将正常贷款样本定义为正样本, 违约样本定义为负样本, 混淆矩阵定义如表 2.

表 2 混淆矩阵定义表

混淆矩阵		预测	
		正类	负类
真实	正类	*TP* (真正例)	*FN* (假反例)
	负类	*FP* (假正例)	*TN* (真反例)

根据混淆矩阵, 定义准确率 A (accuracy)、精准率 P (precision) 与召回率 R (recall), 定义如下:

$$A=(TN+TP)/(FN+FP+TN+TP) \tag{5}$$

$$P=TP/(FP+TP) \tag{6}$$

$$R=TP/(FN+TP) \tag{7}$$

根据上述定义, 可以求得 ROC 曲线, 用以考量针对不平衡样本的预测精度.

3.2 结果分析

经过综合学习粒子群算法对学习速率、正则子参数及贝叶斯套袋控制强度的调节, 经过 120 次迭代, 得到最优参数. 将 CLPSO-CatBoost 与 PSO-CatBoost 的模型误差率进行对比, 二者迭代过程如图 2 所示.

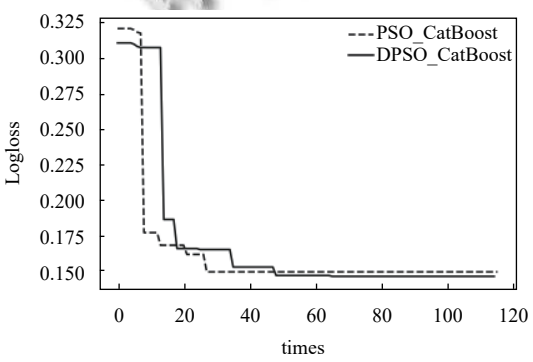

图 2 CLPSO 和 PSO 寻优迭代曲线

PSO 与 CLPSO 算法在优化 CatBoost 模型时, 二者收敛速度相差不大, 均能在 40 次内获得良好的优化效果, 但可以发现, PSO 算法在随后的迭代中陷入了局部最优, 而 CLPSO 算法表现出了更好的局部寻优能力. 由此说明, CLPSO 算法在优化 CatBoost 模型时, 能够跳出局部最优, 所得模型对数损失更小. CLPSO-CatBoost 模型的准确率 90.42% 高于并未进行优化的 CatBoost 模型.

为验证 CLPSO-CatBoost 模型精确度, 实验引入参数优化前的 CatBoost、SVM 和 XGBoost 模型进行训练. 测试集精度如表 3 所示. 根据结果可知, CLPSO-CatBoost 模型在准确率、精准率等方面均有着出色的表现. 相较其他常用模型, CLPSO 在精准率上有着些许优势, 而信用贷款风控因其金融属性对精度有着较高的要求.

表 3 模型性能评价表 (单位: %)

模型	Accuracy	Precision	Recall	AUC
CLPSO-CatBoost	90.42	91.37	89.45	92.58
CatBoost	88.44	90.68	89.32	89.90
XGBoost	88.94	87.47	88.04	85.68
SVM	72.43	74.56	70.41	77.11

根据实验数据绘制各模型 ROC 曲线, 所得曲线如图 3 所示.

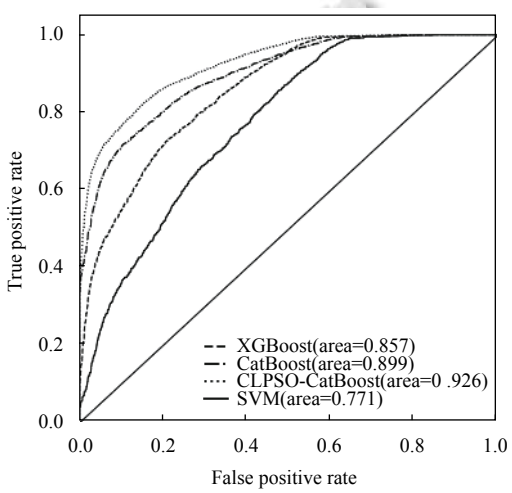

图 3 模型 ROC 曲线

由图 3 可知, CLPSO-CatBoost 模型在实验环境下, 有着很好的表现能力.

该模型作为贷款风险预测模型, 实际中贷款申请数量庞大, 模型性能的提升意味着将会更能判别贷款申请是否具有风险, 从而降低金融公司的坏账率. 车贷

的借款金额较大且订单数量多, 坏账对金融公司造成的影响较为严重. 因此, 模型性能的提升意味着能够有效的减少坏账损失, 进而维护了金融稳定性, 减少发生信用危机的可能性.

4 结论与展望

本文针对金融机构贷款风险预测问题, 通过国内某金融公司的车辆抵押贷款数据, 从数据采样、特征工程及分类算法方面等方面做了一系列工作, 得出如下结论:

1) 相比于传统分类模型. CLPSO-CatBoost 模型在准确率、错误率、召回率和 AUC 曲线上, 都获得了提升. 面对不平衡数据集, 该模型有着出色的少数类识别率, 具有很高的应用价值.

2) CLPSO 作为粒子群算法的一个分支, 通过改进例子学习方式在一定程度上避免了粒子群算法容易陷入局部收敛的问题, 提升了全局学习能力. 在本实验场景下, CLPSO 算法表现出优于 PSO, CLPSO-CatBoost 模型能够有效地提升模型精度.

综上所述, 相比 SVM 等常用的信用风险评估方法, CLPSO-CatBoost 模型能够更有效预测贷款风险.

参考文献

1 Zhu Y, Xie C, Wang GJ, *et al*. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. Neural Computing and Applications, 2017, 28(1): 41–50.

2 Fernandes GB, Artes R. Spatial dependence in credit risk and its improvement in credit scoring. European Journal of Operational Research, 2016, 249(2): 517–524. [doi: 10.1016/j.ejor.2015.07.013]

3 梁琪. 企业信用风险的主成分判别模型及其实证研究. 财经研究, 2003, 29(5): 52–57. [doi: 10.3969/j.issn.1001-9952.2003.05.009]

4 萧超武, 蔡文学, 黄晓宇, 等. 基于随机森林的个人信用评估模型研究及实证分析. 管理现代化, 2014, 34(6): 111–113. [doi: 10.3969/j.issn.1003-1154.2014.06.038]

5 郭春桃. 基于组合模型的个人信用评估研究 [硕士学位论文]. 天津: 天津商业大学, 2019.

6 Dorogush AV, Ershov V, Gulin A. CatBoost: Gradient boosting with categorical features support. arXiv: 1810.11363, 2018.

7 周驰, 高海兵, 高亮, 等. 粒子群优化算法. 计算机应用研究, 2003, 20(12): 7–11. [doi: 10.3969/j.issn.1001-3695.2003.12.003]

8 Liang JJ, Qin AK, Suganthan PN, *et al*. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. IEEE Transactions on Evolutionary Computation, 2006, 10(3): 281–295. [doi: 10.1109/TEVC.2005.857610]

9 Katari V, Malireddi S, Bendapudi SKS, *et al*. Adaptive nonlinear system identification using comprehensive learning PSO. Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing. St. Julians, Malta. 2008. 434–439.

10 陈希, 王斌, 喻敏, 等. 基于 CLPSO 优化 LSSVM 的风数据缺失部分插补. 可再生能源, 2016, 34(6): 878–883.

11 Prokhorenkova L, Gusev G, Vorobev A, *et al*. CatBoost: Unbiased boosting with categorical features. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2018. 6639–6649.

12 蔡文学, 罗永豪, 张冠湘, 等. 基于 GBDT 与 Logistic 回归融合的个人信贷风险评估模型及实证分析. 管理现代化, 2017, 37(2): 1–4. [doi: 10.3969/j.issn.1003-1154.2017.02.001]

13 姜雪莹, 秦进. 基于群决策的 P2P 借贷信用风险评估. 计算机系统应用, 2019, 28(5): 226–231. [doi: 10.15888/j.cnki.csa.006901]

14 郭海山, 高波涌, 陆慧娟. 基于 Boruta-PSO-SVM 的股票收益率研究. 传感器与微系统, 2018, 37(3): 51–53, 57.

15 张昊, 纪宏超, 张红宇. XGBoost 算法在电子商务商品推荐中的应用. 物联网技术, 2017, 7(2): 102–104.

16 Vovk V. The fundamental nature of the log loss function. In: Beklemishev LD, Blass A, Dershowitz N, *et al*., eds. Fields of Logic and Computation II. Cham: Springer, 2015. 307–318.