计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):224−231 [doi: 10.15888/j.cnki.csa.008959]

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

基于特征优化和 Boosting 算法的个人信用预测①

常三强, 周垂日

(中国科学技术大学 管理学院, 合肥 230026)
通信作者: 常三强, E-mail: csq1@mail.ustc.edu.cn

摘　要: 随着互联网金融和电子支付业务的高速增长, 由此引发的个人信用问题也呈现与日俱增的态势. 个人信用预测本质上是不平衡的序列二分类问题, 这类问题的数据样本规模大、维度高、数据分布极不平衡. 为了高效区分申请者的信用情况, 本文提出一种基于特征优化和集成学习的个人信用预测方法 (PL-SmoteBoost). 该方法在 Boosting 集成框架下构建个人信用预测模型, 首先利用 Pearson 相关系数对数据进行初始化分析, 剔除冗余数据; 通过 Lasso 选取部分特征来减少数据维度, 降低高维风险; 通过 SMOTE 过采样方法对降维数据的少数类进行线性插值, 以解决类不平衡问题; 最后为了验证算法有效性, 以常用的处理二分类问题的算法作为对比方法, 采用从 Kaggle 和微软开放数据库下载的高纬度不平衡数据集对算法进行测试, 以 AUC 作为算法的评价指标, 利用统计检验手段对实验结果进行分析. 结果表明, 相对于其他算法, 本文提出的 PL-SmoteBoost 算法具有显著优势.

关键词: 个人信用; SMOTE; 集成学习; 特征优化

引用格式: 常三强,周垂日.基于特征优化和 Boosting 算法的个人信用预测.计算机系统应用,2023,32(3):224–231. http://www.c-s-a.org.cn/1003-3254/8959.html

Personal Credit Prediction Based on Feature Optimization and Boosting Algorithm

CHANG San-Qiang, ZHOU Chui-Ri

(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: With the rapid growth of Internet finance and electronic payment business, resulting personal credit problems are also increasing. Personal credit prediction is essentially an imbalanced binary sequence classification issue. Such an issue is faced with a large size and high dimension of data samples and extremely imbalanced data distribution. To effectively distinguish the credit situation of applicants, this study proposes a personal credit prediction method based on feature optimization and ensemble learning (PL-SmoteBoost). This method involves the construction of a personal credit prediction model within the boosting ensemble framework. Specifically, data initialization analysis with the Pearson correlation coefficient is conducted to eliminate redundant data; some features are selected with the least absolute shrinkage and selection operator (Lasso) to reduce data dimension and thereby lower high dimensional risks; linear interpolation among the minority classes in the dimension-reduced data is carried out by SMOTE oversampling to solve the class imbalance problem; finally, to verify the effectiveness of the proposed algorithm, this study takes the algorithms commonly used to deal with binary classification issues as comparison methods and tests the algorithms with the high dimensional imbalance datasets downloaded from the open databases of Kaggle and Microsoft. With the area under the curve (AUC) as the algorithm evaluation index, the test results are analyzed by the statistical test method. The results show that the proposed PL-SmoteBoost algorithm has significant advantages over other algorithms.

Key words: personal credit; SMOTE; ensemble learning; feature optimization

① 基金项目: 国家自然科学基金面上项目 (72071188)
收稿时间: 2022-07-26; 修改时间: 2022-08-26; 采用时间: 2022-09-01; csa 在线出版时间: 2022-10-28
CNKI 网络首发时间: 2022-11-15

互联网及移动支付等技术的蓬勃发展给传统金融业带来了广泛影响, 网络借贷就是其催生的产物. 一方面, 由于网络借贷具备方便、灵活的融资属性, 越来越多的人选择其作为融资渠道. 另一方面, 由于网络借贷用户的信用水平难以准确识别, 这导致很高的违约率, 严重阻碍了网贷平台的健康发展. 因此, 构建一套广泛适用的个人信用预测模型对网络借贷的风险控制以及网贷平台的良性发展具有深远的意义.

近年来, 在信用风险预测模型的研究中应用了很多基于机器学习和统计学的方法, 例如支持向量机、神经网络、随机森林等方法. 国内外学者对此进行了大量探索研究, 但是在信用风险评估场景下, 样本中出现贷款违约的比例很低, 多数类样本和少数类样本比例相差极大. 使用这样的样本对模型进行训练, 严重影响了模型的预测能力[1]. 针对此类数据集的预测问题, 目前学界主要集中在数据和算法两个层面.

(1) 数据层面一般采用的方法是用欠采样或过采样方法对数据进行处理. Chen 等[2] 在集成学习方法中引入参数扰动并将其与欠采样方法融合构建预测模型, 一定程度上解决了随机欠采样带来的信息遗失问题, 但是在处理多数类和少数类样本极不平衡的数据集时预测效果不够理想. Chawla 等[3] 提出了 SMOTE 算法, 该算法通过人工合成少数类样本使数据集样本分布平衡, 减少了出现过拟合的概率. Phetlasy 等[4] 使用 SMOTE 方法对少数类样本进行过采样处理, 并提出一种组合多个分类器的方法, 该方法将未被正确分类的流量数据送入后续的分类器重新分类, 从而提高预测的灵敏度和准确率. 田臣等[5] 融合了随机森林算法和少数类过采样法构建信用预测模型, 预测效果较随机森林算法和朴素贝叶斯算法相比更好. 针对数据维度过高的问题, Cai 等[6] 使用 Pearson 相关系数来衡量特征变量间的线性相关关系, 从而将相关性较小的特征变量进行筛除, Tibshirani[7] 提出了 Lasso 变量选择方法. Lasso 方法增加 $L1$ 范式函数作为惩罚项来压缩变量系数, 将相关性弱的变量系数压缩为 0, 从而实现特征压缩和对应参数的估计.

(2) 算法层面. 传统分类算法在处理正负样本类极不平衡的数据集时表现较弱, 为了解决这一问题, 学者们将损失函数或错误率引入模型中, 代表方法包括代价敏感学习、集成学习算法等[8]. 代价敏感学习[9] 通过增加少数类样本分类错误的惩罚代价, 优化目标函数从而提升模型的分类准确度. 而集成学习算法[10] 集合了多个基分类器, 降低单个分类器对正负样本类比例相差很大的数据集进行预测时出现的偏差, 提升整体模型的分类准确度. 目前广泛使用的是融合数据采样方法与分类算法. Sun 等[11] 提出一种对不平衡数据集进行分类的方法, 该方法将不平衡数据集分解为数个平衡的数据子集, 然后训练每个子集从而获得基分类器. 王俊红等[12] 提出一种融合了代价敏感学习算法和欠采样法的预测方法, 提升了处理样本不平衡数据集时的分类效果.

随着大数据时代的全面到来, 用户特征呈现高维度稀疏发展趋势, 从而导致分类模型难以正确区分个人信用数据的多数类和少数类. 同时高维度数据会导致维度灾难, 使算法的计算开销随着数据维度的上升出现指数增长. 因此, 维度灾难和样本类不平衡这两个难题[13], 严重影响了个人信用评价问题.

综上所述, 本文针对高维度不平衡数据集进行预测的问题, 提出一种基于特征优化和集成学习的个人信用预测方法 (PL-SmoteBoost). 主要贡献如下: 1) 利用 Pearson 对数据进行初始化分析, 剔除冗余数据; 通过 Lasso 选取部分特征来减少数据维度, 降低高维风险. 2) 通过 SMOTE 过采样方法对降维数据的少数类进行线性插值, 以解决类不平衡问题. 3) 构建基于集成学习的个人信用评估模型, 以常用的处理二分类问题的算法作为对比方法, 采用从 Kaggle 和微软开放数据库上下载的数据集对算法进行测试, 证明该方法的有效性.

1 特征优化的理论与方法

1.1 Pearson 相关系数

皮尔森 (Pearson) 系数[14] 具有去中心化、归一化等特点, 在反应目标值与特征值的相关性上有着出色表现, 因此广泛应用于衡量特征的线性相关程度上, 其公式如下:

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{1}$$

其中, 变量x为特征值, 变量y为信用分, Pearson 相关系数即为上述 2 个变量的协方差除以其标准差的乘积. 其中系数$r(x,y)$的取值范围在−1 至 1 之间, 值越接近

1 或−1 则表示正负相关性越强, 值越接近 0 表示相关性越弱. 皮尔森系数在保留偏好特征方面表现良好, 能较好反映特征间的线性相关性, 因此适合用于信用评价中特征的筛选.

1.2 Lasso 算法

Lasso 算法是一种基于线性回归模型的特征筛选办法, 通过对变量进行筛选和压缩来减少特征维度, 可以有效防止出现过拟合问题. Lasso 算法使用 $L1$ 范式构造惩罚函数, 在多元线性回归误差平方和最小的基础上, 对回归系数增加惩罚函数, 将与模型结果相关性较小的变量回归系数压缩至 0, 从而删除这些特征变量, 达到减少特征维度的目的[15].

Lasso 特征选择方法即为残差平方和最小化加上 $L1$ 范式的惩罚项, 公式如下:

$$\min\sum e_i^2 + L1\text{范式} = \min\sum (x_i - \hat{x}_i)^2 + \lambda\sum_{u=1}^{k}\left|\hat{\beta}_u\right| \quad (2)$$

其中, x_i表示个体i的变量x的实际数值; $\hat{x}_i$表示对个体i的变量x的估计值; $e_i = x_i - \hat{x}_i$, 代表了估计值与实际值之间的差值; $\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k)$ 为回归系数. λ是预先设定好的非负数, 其大小决定了算法的选择能力, 如果λ过大则会导致无法删除任何特征, 通过将λ设定为一个合适的低值, 可以将与类相关性较低的特征变量系数降为 0, 从而删除这些特征, 突出与类相关性强的特征, 实现对数据集的特征筛选.

1.3 SMOTE 算法

SMOTE 合成少数类过采样技术是一种基于随机过采样法的改进方法. 它基于“线性插值”来合成新的少数类样本[16], 以每个少数类样本的 K 个最近邻样本为参照, 随机的选择数个邻近点进行插值, 从而合成新的少数类样本. 因为特征空间上临近的样本特征具有相似性, 据此合成的新样本与老样本间特征也相似.

样本不平衡问题指正样本数量远高于负样本数量, 通过 SMOTE 技术合成少数类样本, 从而有效解决数据集的样本不平衡问题. 合成少数类样本的公式如下:

$$x_n = P + \lambda \times (K_i - P) \quad (3)$$

其中, x_n为新构造的少数类样本; P为少数类中的每一个样本; λ为区间 (0, 1) 内的随机数; K_i为数据样本P中最近邻样本中的第 n 个样本; i=1, 2, …,k; 通过 SMOTE 技术构造少数类样本, 可以有效缓解数据集类不平衡问题.

SMOTE 算法的合成实例如图 1 所示, 首先依次选择数据集中的每个少数类样本P; 然后从P的最近邻样本中选择 n 个, n 一般为 5; 依次在选择的 n 个最近邻样本中与少数类样本P使用式 (3) 进行线性插值, 从而获得新合成的少数类样本x_n, 通过合成少数类样本, 使得数据集的正负样本数量趋于均衡, 提升模型预测准确度.

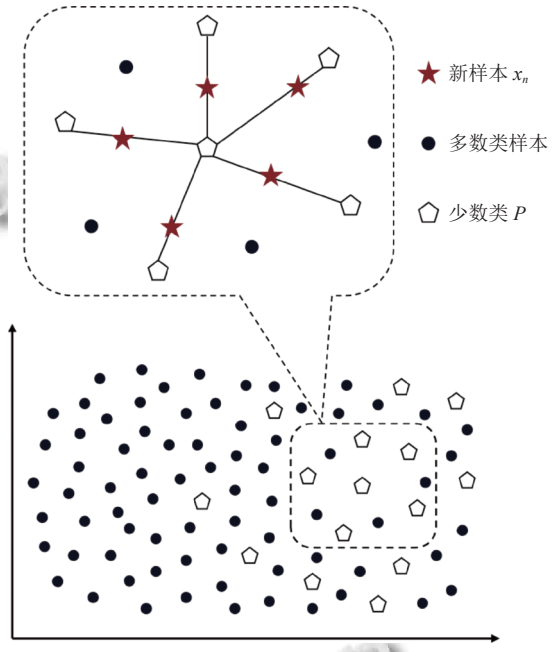

图 1 SMOTE 算法实例

1.4 Boosting 集成学习算法

XGBoost 是一种基于传统梯度提升树 (GBDT) 进行改进的算法, 它使用预排序算法对树模型进行拟合, 在树生长的过程中遍历每一个切分点, 从而找到最优切分点对数据做叶子节点的切分, 获得各个叶子节点的分数, 将其加和得出样本的预测值. 并且 XGBoost 内置处理缺失值的规则, 支持并行计算, 灵活性很高, 能在消耗更少内存的情况下提供更快的运算速度[17].

LightGBM 算法基于 XGBoost 做了进一步改进. LightGBM 使用基于叶子节点分割的树生长策略以及直方图算法, 大幅度降低了计算和内存消耗, 因而可以在维持模型精度的前提下提升训练速度[18]. 并且 LightGBM 提升了树的最大深度限制, 一定程度可以规避按叶子节点分割所导致的过拟合问题.

CatBoost 是以对称树为基学习器的算法, 在训练过程中处理类别特征时使用 Target-based 方式, 因而

CatBoost 适宜用来处理大部分特征都是类别特征的数据集. 另外 CatBoost 降低了对广泛的超参数优化的依赖, 一般情况下使用默认参数就能获得良好的效果, 有助于降低预测所需的时间, 规避过拟合问题[19].

2 基于 PL-SmoteBoost 个人信用预测方法

为了解决面向高纬度、不均衡数据环境下的信用分预测问题, 提出一种基于特征优化和集成学习的个人信用预测方法 (PL-SmoteBoost). 该模型分为 3 个部分, 各部分介绍如下.

(1) 数据清洗部分. 数据集的数据来源多种多样, 例如人工录入、用户填写、爬虫爬取等, 因为用户填写不认真、爬虫程序不够智能、人工录入失误、存储设备故障等原因, 数据难以避免的会产生大量缺失值和异常值, 例如一条数据只有名字没有其他信息, 或者年龄一栏错填了性别等问题. 这使得数据集的预测能力大打折扣. 因此, 为了提升预测水平, 将获取的数据集进行缺失值填充、异常值删除等操作, 提高模型训练能力.

(2) 模型构建部分. 分为数据模块和训练模块, 数据模块是通过统计学和机器学习等方法将原始的数据转换成可参与模型计算的形式. 训练模块采用集成学习中 Boosting 算法对模型进行训练.

(3) 模型评估部分. 输出基于 PL-SmoteBoost 的个人信用评估方法的结果, 并以 AUC 值作为算法的评价指标.

2.1 算法流程

基于特征优化和集成学习的个人信用预测方法 (PL-SmoteBoost) 的流程图如图 2 所示.

基于 PL-SmoteBoost 的个人信用预测模型的具体步骤如下.

第 1 步. 数据预处理. 对获取到的数据集 A、B 进行缺失值填充、异常值删除等操作, 得到数据集 A_0、B_0, 使数据处于同一量纲下, 提升数据表达能力.

第 2 步. 使用 Pearson 系数分析特征与信用之间的关联性, 采用 Lasso 提取重要程度高的特征, 进行特征选择, 减少数据维度, 降低高纬风险.

第 3 步. 采用 SMOTE 过采样技术对不平衡数据进行衍生处理, 输出均衡数据集 A_1、B_1.

第 4 步. 模型训练. 将经过特征处理后的数据 A_1、B_1, 分别代入 Boosting 集成学习 LightGBM、XGBoost、CatBoost 算法进行模型训练, 得出预测结果.

第 5 步. 评分结果. 输出基于 PL-SmoteBoost 的个人信用评分模型的预测结果, 以常用的处理不均衡分类问题的算法作为对比方法进行测试, 以 AUC 值作为算法的评价指标, 利用统计检验手段对实验结果进行分析得出结论, 基于 PL-SmoteBoost 的个人信用评分模型训练效果好, 适合应用于信用评分预测.

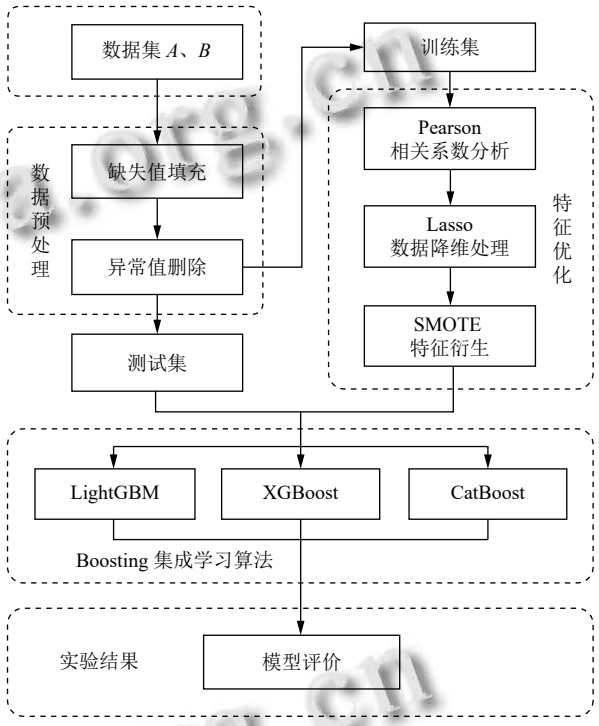

图 2 PL-SmoteBoost 个人信用预测流程图

2.2 评价指标

评价一个分类器的好坏有许多标准, 一般情况下准确率高的分类器效果就好. 但是在多数类和少数类样本比例差别很大的情况下, 准确率高并不一定意味着分类器性能好. 例如对用户的信用情况进行预测, 因为信用良好的用户数量远比信用违约的用户数量高, 所以在大部分情况下, 任意模型预测客户的信用良好, 准确率都能达到 90% 以上, 因此使用准确度对分类器的性能进行评价并不充分.

因此本文采用 AUC 值对分类器进行评价. ROC 曲线是以真正率为坐标轴纵轴、假正率为坐标轴横轴的曲线, AUC 值为 ROC 曲线和 X 轴围成的面积. ROC 曲线越贴近坐标轴左上角则表明模型的性能越好[20], 即 AUC 值越大越好. AUC 值在样本数据失衡的情况下仍能较好地反应模型预测能力, 因此为了量化模型

的好坏, 用 AUC 值来评价分类器的性能.

3 实验分析

3.1 数据清洗

本文使用两个数据集进行对照实验, 以增强模型证明力度. 数据为脱敏后的数据, 数据来源为在 Kaggle 下载的数据集 *A* 和微软开放数据库下载的数据集 *B*.

本部分将以 Kaggle 数据集 *A* 为例详细描述对数据集的处理过程. 数据集 *A* 数据特征多达 56 个, 共分为 5 类, 其中包括用户基本信息、借贷行为、活跃行为、用户消费行为、违约情况.

(1) 缺失值处理

数据集 *A* 原始的 56 个特征中: 有 42 个特征含有缺失值, 有 12 个特征的缺失值比例在 30% 以上, 缺失值比例过大的特征直接整列删除, 4 个特征采用固定值填充. 以工作年限这列特征为例, 统计可知其中有 3 376 条缺失值, 原始数据集中用“n/a”表示, 本步采用其他样本的均值对其进行填充.

(2) 异常值检测

异常值是数据集中的非正常值, 其可能是不正确的“脏数据”, 也可能是正确的异常数据, 需要具体分析后, 对部分特征采用均值填充进行处理, 部分无法处理的数据做删除处理.

对缺失值、异常值处理后保留了 29 个特征变量, 共计 6 889 个样本. 其中, 违约样本数量为 682 个, 未违约样本数量为 6 207 个. 本文建立如表 1 所示数据特征指标, 对个人信用风险进行识别.

3.2 特征优化

(1) Pearson 相关系数分析

如图 3 所示, 为了降低数据集特征维度, 提高数据分析的准确性和高效性. 将经过预处理的特征变量输入模型算法中, 计算各个特征与信用预测之间的相关系数, 将其进行排序并输出为一个特征值与信用预测相关性的柱状图, 从图中可以发现部分特征与信用预测相关性较低, 将这些特征删除, 保留剩下的与信用预测相关性较高的特征变量. 本步骤剔除的是登录频率和用户使用 APP 时间、交往圈人数、社交频率 4 个低相关特征, 保留剩下的 25 个特征.

(2) Lasso 特征选择

Lasso 算法通过增加 *L*1 范式函数作为惩罚项, 将绝对值小于阈值的特征回归系数压缩至 0, 即将这些特征变量对于分类结果的贡献忽略, 从而可以剔除这些变量. 在 Lasso 算法中, λ的大小决定了算法的选择能力, 因此本步骤中微调其大小, 从中选择最合适的值. Lasso 中参数λ微调结果如表 2 所示.

表 1 数据特征指标及说明

类别	序号	指标名称	说明
基本信息	A01	用户年龄	用户出生至今的年龄
	A02	用户性别	1=男性, 0=女性
	A03	婚姻状况	1=已婚, 0=未婚
	A04	受教育情况	用户受教育的程度
	A05	使用手机品牌	用户当前使用手机的品牌
	A06	是否就业	1=已就业, 0=未就业
	A07	工作年限	用户的工龄
	A08	住房性质	1=自有产权, 0=租房
借贷行为	A09	贷款笔数	用户当前贷款笔数
	A10	贷款金额	用户当前贷款金融总数
	A11	贷款周期	用户借款至还款一个周期的时间
	A12	授信额度使用率	用户使用平台授信额度的比率
	A13	提前还款	1=有提前还款, 0=无提前还款
	A14	申贷次数	用户向平台申请贷款的次数
	A15	拒贷次数	平台拒绝用户申请贷款的次数
	A16	平台外申贷	用户是否在其他平台具有贷款
活跃行为	A17	登录频率	用户登录APP的频率
	A18	使用时间	用户使用APP的时间
	A19	通话频率	用户使用运营商通话的频率
	A20	交往圈人数	用户当月通话交往圈人数
	A21	社交频率	用户通过APP与好友互动的次数
	A22	营销响应行为	用户响应平台推送广告的行为
消费行为	A23	消费笔数	用户当月消费次数
	A24	消费金额	用户当月消费总金额
	A25	消费类目	用户当月消费总类目
	A26	信用卡账单	用户当月信用卡账单数目
	A27	电商流水账单	用户当月电商流水账单数目
	A28	用户月均话费	用户每月平均使用话费的数额
违约情况	A29	是否有逾期贷款	1=存在逾期贷款, 0=正常贷款

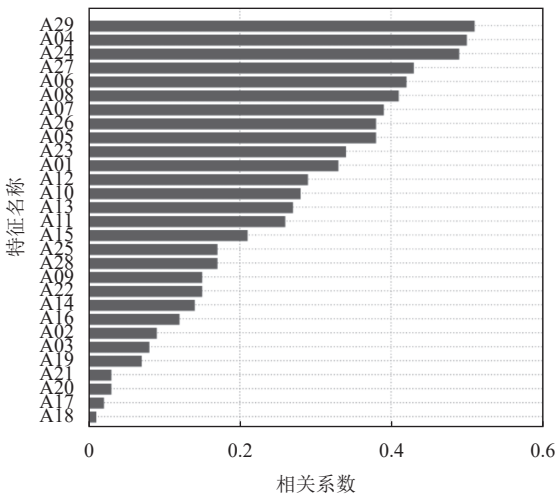

图 3 特征与信用分相关性

表 2 不同λ值变量数对比

变量数	5	15	22	25
λ	0.01	0.005	0.002 5	0.000 1

当λ为 0.01 时, 模型存在欠拟合的情况, 25 个特征变量中只保留了 5 个. 经过多次测试训练集与测试集得分, 在λ为 0.005–0.002 5 之间时训练集与测试集分值相差较低. 当λ降至 0.000 1 时, 训练集与测试集得分差值较高, 存在过拟合现象. 因此本文选择λ=0.0025 进行实验, 此时样本保留 22 个.

(3) SMOTE 过采样

如图 4 所示, 本文采用的用户数据中正负样本不均衡, 正样本数目远大于负样本. 而大部分分类器会基于阈值输出结果, 如大于某值的为正例, 反之则为反例. 在样本数据不均衡时, 预先设定的阈值会导致模型输出结果倾向于数据多的类别. 为了解决这一问题, 本文采用 SMOTE 过采样技术合成少数类样本, 让正负样本数目一样多.

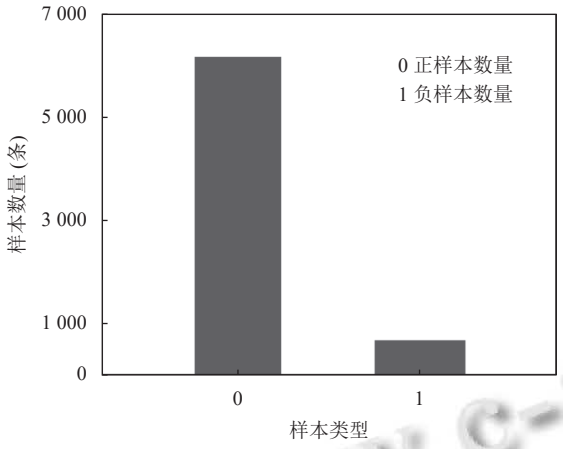

图 4 正负样本数量对比

SMOTE 算法步骤如算法 1.

算法 1. SMOTE 算法

输入: 少数类样本P, 向上采样倍率N, 样本近邻数k
输出: 新合成的少数类样本x_n

步骤 1. 计算得到少数类样本x_n的k个近邻样本
步骤 2. 在k个近邻样本中随机选择一个为K_i
步骤 3. 生成一个 0 到 1 之间的随机数λ
步骤 4. 按照向上采样倍率N重复步骤 2–3
步骤 5. 合成新的少数类样本x_n
步骤 6. 对新合成的少数类样本x_n做分类
步骤 7. 输出分类后的结果

3.3 实验评估

经过数据清洗和特征优化后, 本文获得了数据集A_1、B_1. 为了验证 PL-SmoteBoost 模型对个人信用评分的预测能力, 使用 Python 3.9.0 开发环境, 进行如下两组实验.

实验 1. 本轮实验为了验证特征优化对模型预测效果的影响, 实验采用仅经过数据清洗的数据集A_0、B_0以及特征优化后的数据集A_1、B_1进行对比, 分别使用 SVM、LR、RF、BP 四种算法进行实验对比分析.

表 3 为特征优化前后 AUC 值对比. 从实验结果可以看出, A、B 两个数据集的 4 种模型的训练子集在经过特征优化后, AUC 值都有所提升, 提升的数值约在 7.2%–9.6% 之间, 说明本文特征优化处理后的数据对提升模型准确率有显著的效果, 其中 RF 算法的优化效果最好.

表 3 特征优化前后 AUC 值对比

模型	特征优化前		特征优化后	
	A_0	B_0	A_1	B_1
SVM	0.591	0.603	0.683	0.677
LR	0.651	0.626	0.723	0.709
RF	0.662	0.657	0.758	0.751
BP	0.654	0.649	0.729	0.734

实验 2. 本轮实验为了体现 Boosting 集成学习算法对个人信用预测的作用, 引入基于 Boosting 集成学习中的 XGBoost、CatBoost、LightGBM 和实验 1 中预测效果最优的 RF 算法进行对比, 采用五折交叉算法对数据集进行训练来减少随机性对分类结果的影响. 实验结果如表 4、表 5 所示.

表 4 基于 Kaggle 数据集 A 的 4 种模型 AUC 值对比

模型	Fold1	Fold2	Fold3	Fold4	Fold5	平均
XGBoost	0.862	0.842	0.916	0.891	0.873	0.877
CatBoost	0.763	0.707	0.734	0.837	0.852	0.779
LightGBM	0.698	0.787	0.801	0.743	0.721	0.750
RF	0.724	0.681	0.643	0.781	0.731	0.712

表 5 基于微软数据集 B 的 4 种模型 AUC 值对比

模型	Fold1	Fold2	Fold3	Fold4	Fold5	平均
XGBoost	0.848	0.905	0.874	0.816	0.875	0.864
CatBoost	0.743	0.857	0.825	0.785	0.879	0.818
LightGBM	0.825	0.692	0.849	0.792	0.738	0.779
RF	0.729	0.658	0.736	0.783	0.697	0.721

从表 4、表 5 可以看出, 在 4 个评估模型中, 基于 Boosting 的 3 个模型效果都比 RF 好. 其中, XGBoost 模型训练效果最好, A、B 两组数据集的 AUC 平均值

分别为 0.877 和 0.864, 比 RF 模型的 AUC 平均值分别高出 16.5% 和 14.3%, LightGBM 的训练效果在 3 组 Boosting 算法中最差, 但其 AUC 平均值也比 RF 模型的 AUC 平均值分别高出 3.8% 和 5.8%.

为了能够更直观的对比 4 种模型的训练效果, 本文使用 ROC 曲线对比图来呈现, 如图 5、图 6 所示, ROC 曲线是一条在不同阈值下分类结果的假正率和真正率所构成的曲线, 其中坐标轴横轴为假正率, 坐标轴纵轴为真正率, 曲线下的面积为 AUC 值, 即曲线越靠近左上方模型分类效果越好.

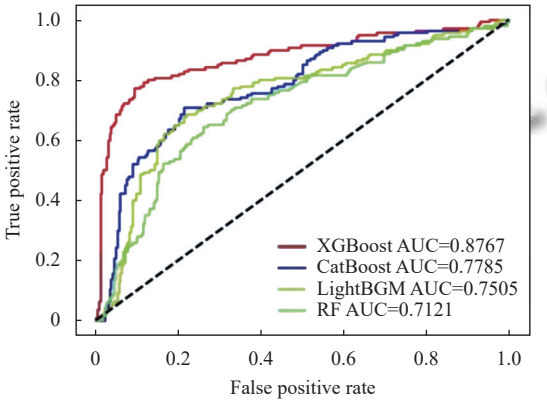

图 5 基于 Kaggle 数据集 *A* 的 ROC 曲线对比

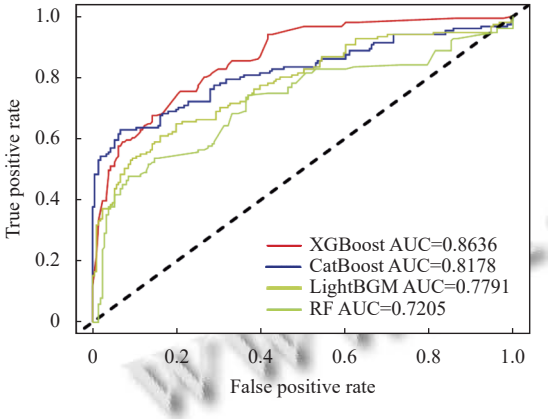

图 6 基于微软数据集 *B* 的 ROC 曲线对比

从图 5、图 6 中 4 种模型的 ROC 曲线可以看出, XGBoost 都在图中最上方, 训练效果最优; 而 RF 的 ROC 曲线都在最下方, 结果相比 Boosting 算法中的 3 个模型要差.

4 总结与展望

为了解决面向高纬度、不均衡数据环境下的信用分预测问题, 本文提出一种基于特征优化和集成学习的个人信用预测方法 (PL-SmoteBoost). 该方法充分考虑了数据高纬度、不均衡等特点, 选取 Pearson 相关系数、Lasso、SMOTE 等算法进行数据特征的优化, 考虑个人信用评分的特殊性, 采用对噪声敏感的 Boosting 集成学习算法对数据进行训练. 选用 Kaggle 和微软开放数据库上公开的数据集进行评估, 使用 AUC 值作为评估指标, 实验结果表明, PL-SmoteBoost 算法性能优于其他算法, 适用于个人信用评分.

不足及未来展望: 本文采用 SMOTE 过采样方法, 虽然解决了数据不平衡问题, 但是还是存在一些问题: 比如合成样本的质量问题、模糊类边界问题、少数类分布问题. 这些都会影响模型最终的预测结果, 所以选取更加合适的数据平衡方法是需要进一步探讨的问题.

参考文献

1 王重仁, 韩冬梅. 基于超参数优化和集成学习的互联网信贷个人信用评估. 统计与决策, 2019, 35(1): 87–91.
2 Chen QW, Wang W, Ma D, *et al*. Class-imbalance credit scoring using Ext-GBDT ensemble. Application Research of Computers, 2018, 35(2): 421–427.
3 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.
4 Phetlasy S, Ohzahata S, Wu C, *et al*. Applying SMOTE for a sequential classifiers combination method to improve the performance of intrusion detection system. Proceedings of 2019 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress. Fukuoka: IEEE, 2019. 255–258.
5 田臣, 周丽娟. 基于带多数类权重的少数类过采样技术和随机森林的信用评估方法. 计算机应用, 2019, 39(6): 1707–1712. [doi: 10.11772/j.issn.1001-9081.2018102180]
6 Cai JC, Xu K, Zhu YH, *et al*. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied Energy, 2020, 262: 114566. [doi: 10.1016/j.apenergy.2020.114566]
7 Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267–288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
8 王忠震, 黄勃, 方志军, 等. 改进 SMOTE 的不平衡数据集

成分类算法. 计算机应用, 2019, 39(9): 2591–2596. [doi: 10.11772/j.issn.1001-9081.2019030531]
9 吴雨茜, 王俊丽, 杨丽, 等. 代价敏感深度学习方法研究综述. 计算机科学, 2019, 46(5): 1–12. [doi: 10.11896/j.issn.1002-137X.2019.05.001]
10 赵楠, 张小芳, 张利军. 不平衡数据分类研究综述. 计算机科学, 2018, 45(6A): 22–27, 57. [doi: 10.11896/j.issn.1002-137X.2018.Z6.004]
11 Sun ZB, Song QB, Zhu XY, *et al*. A novel ensemble method for classifying imbalanced data. Pattern Recognition, 2015, 48(5): 1623–1637. [doi: 10.1016/j.patcog.2014.11.014]
12 王俊红, 闫家荣. 基于欠采样和代价敏感的不平衡数据分类算法. 计算机应用, 2021, 41(1): 48–52.
13 杨平安, 林亚平, 祝团飞. AdaBoostRS: 高维不平衡数据学习的集成整合. 计算机科学, 2019, 46(12): 8–12. [doi: 10.11896/jsjkx.180901813]
14 闫政旭, 秦超, 宋刚. 基于 Pearson 特征选择的随机森林模型股票价格预测. 计算机工程与应用, 2021, 57(15): 286–296. [doi: 10.3778/j.issn.1002-8331.2011-0419]
15 许赟娟, 罗幼喜. 基于变量聚类的主成分 Lasso 降维算法与模拟. 统计与决策, 2021, 37(4): 31–36.
16 石洪波, 陈雨文, 陈鑫. SMOTE 过采样及其改进算法研究综述. 智能系统学报, 2019, 14(6): 1073–1083. [doi: 10.11992/tis.201906052]
17 李欣, 俞卫琴. 基于改进 GS-XGBoost 的个人信用评估. 计算机系统应用, 2020, 29(11): 145–150. [doi: 10.15888/j.cnki.csa.007624]
18 任师攀, 彭一宁. 基于软投票融合模型的消费信贷违约风险评估研究. 金融理论与实践, 2020, (4): 77–83. [doi: 10.3969/j.issn.1003-4625.2020.04.010]
19 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. [doi: 10.15888/j.cnki.csa.007866]
20 迟国泰, 张亚京, 石宝峰. 基于 Probit 回归的小企业债信评级模型及实证. 管理科学学报, 2016, 19(6): 136–156. [doi: 10.3969/j.issn.1007-9807.2016.06.010]

(校对责编: 牛欣悦)