Finance 金融, 2020, 10(6), 548-559
Published Online November 2020 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2020.106057

Hans汉斯

基于BLS-分类器的信用评分研究

李超杰[1]，耿玉峰[2]

[1]交通银行江苏省分行，江苏 南京
[2]东南大学，江苏 南京
Email: lichaojie@bankcomm.com, fycs445006997@163.com

收稿日期：2020年10月29日；录用日期：2020年11月12日；发布日期：2020年11月19日

摘 要

为了有效管理信用风险，金融机构开发出各种定量模型。各种分类方法被用来用于信用评分：区分好的和坏的借款人。但信贷数据集存在不平衡性不进行处理会影响分类器性能，且目前并未用不同分类器对单一信贷数据集进行处理。本文基于BLS算法研究了各类分类器对同一信贷数据集的分类性能问题。首先给出了各分类器的分类原理。接着，基于BLS算法对不平衡信贷数据进行处理，再接着给出了分类器的模型分析，最后，利用各项评估指标对不同分类器性能进行比较。结果显示，随机森林分类器最适合对信贷数据集进行分类，未来可将随机森林作为基分类器开发集成模型进一步提高预测准确性和性能。

关键词

信用评分，BLS算法，分类器，评估

Research on Credit Scoring Based on BLS-Classifications

Chaojie Li[1], Yufeng Geng[2]

[1]Bank of Communications Jiangsu Branch, Nanjing Jiangsu
[2]Southeast University, Nanjing Jiangsu
Email: lichaojie@bankcomm.com, fycs445006997@163.com

Received: Oct. 29th, 2020; accepted: Nov. 12th, 2020; published: Nov. 19th, 2020

Abstract

In order to manage credit risk effectively, financial institutions have developed various quantitative models. Various categories are used for credit scoring: to distinguish good and bad borrowers. However, the unbalance of the credit data set will affect the performance of the classifier, and dif-

文章引用: 李超杰, 耿玉峰. 基于 BLS-分类器的信用评分研究[J]. 金融, 2020, 10(6): 548-559.
DOI: 10.12677/fin.2020.106057

ferent classifiers are not used to process a single credit data set. Based on BLS algorithm, this paper studies the classification performance of various classifiers for the same credit data set. Firstly, the classification principle of each classifier is given. Then, the unbalanced credit data is processed based on THE BLS algorithm, and then the model analysis of the classifier is given. Finally, the performance of different classifiers is compared with each evaluation index. The results show that the random forest classifier is the most suitable for the classification of credit data set, and the integrated model can be developed with the random forest as the base classifier in the future to further improve the accuracy and performance of the prediction.

Keywords

Credit Scoring, Border-Line Smote Algorithm, Classifications, Evaluation

1. 引言

随着互联网金融和电子商务的飞速发展，基于互联网平台的信贷业务迅速崛起，金融机构需要对申请贷款的客户进行选择，筛选出有资格获取贷款的客户以做出金融机构的信用决策[1]。而信用评分，作为评估信贷客户的一种分类技术，可以将借款人划分为两种：信用良好的借款人，信用不好的借款人[2]，如今信用评分已经成为信贷行业一项非常重要的任务，金融机构需要开发他们的信用评分模型，以准确地发现他们的不良借款人，这种有效的信用风险管理对银行的长期生存和整个金融体系的稳定至关重要。

从 20 世纪 60 年代末开始随着信贷行业快速发展，信用评分技术也随之不断发展，许多研究提出了新的模型和算法来提高信用评分的准确性，信用评分模型可以分为两类：基于计量统计的传统模型和基于人工智能技术的新型模型[3]。信用评分中最常见和常用的传统模型是线性判别模型和逻辑回归模型，而线性判别分析中的假设与实际不符：信贷数据中好的借款人和坏的借款人的信贷类别的协方差矩阵不可能相等。而逻辑回归模型，作为预测二分结果的首选技术，可以提供良好的准确性和解释率，被认为是一项合适的技术之一去进行信用评分[4]。随着人工智能的发展，出现了一系列非参数方法：决策树学习，遗传算法和神经网络等[5]，这些方法被作为分类器来对信贷用户进行分类，以最小化训练集上的分类误差。通过开发可靠的信用评分系统，可以降低信用分析的成本，实现更快的决策，降低可能的信用风险[6]。

W. E. Henley 采用 KNN 方法对信用体系的违约问题进行分类，研究了模式识别和非参数统计中的标准技术 k-近邻方法在信用评分问题中的应用[7]。支持向量机(SVM)是 Vapnik 提出的一种统计分类方法[8]。支持向量机是根据统计学习理论和遵循结构风险最小化原则设计的分类算法。SVM 分类器作为超平面的几何表示，可以得出唯一最优的凸优化问题的公式，以及估计未知测试数据的泛化误差上限的可能性。所得分类器即使在高维输入空间和小训练样本条件下也能很好地推广。由于可用的金融违约数据量通常很小，而且质量通常很低，因此支持向量机分类器似乎特别适用违约分类问题[9]。针对信用评分问题，多位学者采用了 svm 分类器进行分类[10] [11]。Kiran 基于朴素贝叶斯算法和 KNN 算法用于信用卡欺诈检测[12]。Okesola 采用贝叶斯分类器在构建银行业信用评分模型，以人口统计和物质指标为输入变量考

察贝叶斯的性能[13]。Yung-Chia Chang 使用决策树过滤短期违约，以产生一个高度准确的模型来区分违约贷款人和守信贷款人[14]。

信贷数据集往往是不平衡的，针对不平衡的数据集进行分类往往会得出不精确的结果。对不平衡数据集，学者们提出了欠采样和过采样技术。Nitesh 提出的 SMOTE 算法[15]是一种过采样技术，它能强有力地平衡数据，已经在各种应用领域得以成功应用。由于边界线上的样本容易被错误分类，韩慧，王文渊提出了一种改进的 SMOTE 算法：BLS (BorderLine-Smote)算法，成功提高了预测精度[16]。BLS 算法已在医学研究、网络安全等领域得到应用[17] [18]。

基于信贷数据的不平衡性，广泛使用的 BLS 算法切合平衡数据集的需求。同时已有研究分别用不同分类器对违约分类的信用评分问题进行分类，但仍没有对各类分类器在同一数据集上进行处理。金融机构的主要关注点之一是一个有效的信用评分模型，它是否提供了良好的预测能力。利用不同分类器对同一样本集进行处理，评估它们的性能，可以给金融机构的信贷客户分类提供一定的价值，同时为信贷分类开发集成深度学习模型提供参考。

本文的主要目标是将 BLS 算法用于信贷行业以提高分类器的预测精度，首先用该算法对客户信用不平衡数据集进行处理后得到平衡数据集，再使用多种机器学习算法：逻辑回归、朴素贝叶斯、随机森林和支持向量机来研究信用评分二分类问题，并对它们的性能和对问题的适用性进行评估和讨论。

2. 相关理论

2.1. BLS 算法

BLS (Border Line-Smote)算法是基于 smote 算法进行改进的一种过采样算法，smote 算法通过随机采样合成新样本，而 BLS 算法仅仅使用边界上的少数类样本，以该样本为基础合成新样本，从而改善样本的类别分布。首先需要将少数类样本进行分类，将样本周围一半以上为多数类样本视为边界上的样本，分类为 Danger 类，产生新样本的过程可用如下式表示：

$$New\left(X_i\right)=X_i+dif_j\times\delta \tag{1}$$

其中 $New\left(X_i\right)$ 表示产生的新样本，X_i 表示被分为 Danger 类中的多数类，dif_j 表示原样本 X_i 的 K 近邻，K 近邻是指沿 x 的 n 维特征空间，其自身之间的欧氏距离和权重最小的 K 个元素，δ 指 $[0,1]$ 之间的随机数。计算 K 近临和随机数的乘积，加到原始样本中，即产生新的少数类样本。

2.2. 逻辑回归模型

逻辑回归模型于 1944 年由 Beckson 提出[19]。对于信用评分问题，逻辑回归模型经常被用来与其他模型进行比较[20] [21]。logistic 回归模型作为预测二分法结果的选择技术常被用于信用评分。逻辑回归是一种预测二分因变量的回归方法。在产生逻辑回归方程时，最大似然比被用来确定变量的统计显著性，逻辑回归在对基于一组预测变量的值来预测特征或结果的存在与否的问题上表现良好。逻辑回归类似于线性回归模型，但适用于因变量是二分的模型。在逻辑回归模型中有 s 个自变量，则逻辑回归模型可由下式表示：

$$P\left\{Y=1\right\}=\frac{1}{1+\mathrm{e}^{-\left(\beta_0+\beta_1x_1+\cdots+\beta_sx_s\right)}} \tag{2}$$

其中 $P\left\{Y=1\right\}$ 表示二分类中样本发生的概率，在信贷问题中表示客户守约的概率，而 $\beta_0,\beta_1,\cdots,\beta_s$ 表示回归系数，在逻辑回归模型中隐含着一个线性模型，如下：

$$\ln\left(\frac{P\left\{Y=1\right\}}{1-P\left\{Y=1\right\}}\right)=\beta_0+\beta_1x_1+\cdots+\beta_sx_s \tag{3}$$

逻辑回归模型可以包含主要效果和交互项。在对一组数据建模的过程中，一个重要的步骤是确定数据中是否存在交互作用和因变量和主要自变量相关的协变量。当两种关联都存在时，自变量和因变量之间的关系被认为是混杂的。检查协变量混杂状态的逻辑回归模型是比较包含和不包含协变量的模型中自变量的估计系数。独立变量估计系数的任何临床重要变化都表明协变量是一个混杂因素，应该包括在模型中，而不管其估计系数的统计显著性如何。一种测试协变量和交互作用的方法是从主效应模型开始，并使用向前选择方法来寻找显著降低似然比测试统计量的交互作用项。

2.3. 朴素贝叶斯法

朴素贝叶斯是基于贝叶斯定理的一种分类方法。通过训练数据集，朴素贝叶斯法可以学习到特征值和目标值之间的联合概率分布 $F(X,Y)$。首先学习先验概率分布 $P\{Y=p_k\},k=1,2,3,\cdots,n$，再学习条件概率分布，如下：

$$P\{X=x\mid Y=p_k\}=P\left\{X^1=x^{(1)},X^2=x^{(2)},\cdots,X^n=x^{(n)}\mid Y=p_k\right\} \tag{4}$$

将先验概率分布与条件概率分布相乘得到联合概率分布，在假设数据集的特征独立的情况下，在对目标值进行分类时，在输入 x 给定的情况下，针对条件概率分布，贝叶斯法假设特征相互独立，过给定的数据来计算后验概率分布 $P\{Y=p_k\mid X=x\}$，求出后验概率最大的输出值 y，故朴素贝叶斯分类器可用下式表示：

$$y=\arg\max\frac{P(Y=p_k)\prod P\left\{X^j=x^{(j)}\mid Y=p_k\right\}}{\sum P(Y=p_k)\sum P\left\{X^j=x^{(j)}\mid Y=p_k\right\}\prod P\left\{X^j=x^{(j)}\mid Y=p_k\right\}} \tag{5}$$

2.4. 决策树和随机森林

决策树，作为一个基本的分类和回归方法，决策树利用给定特征条件下的概率分布可以快速进行分类，通过给定损失函数，设定损失函数最小化的优化问题，从而从机器学习中训练出符合条件的模型。L. Breiman 于 2001 年提出随机森林模型[22]，如今已被作为一种通用的分类和回归方法。随机森林方法结合了几个随机决策树，并通过平均来聚集它们的预测，在变量数量远大于观测值数量的环境中表现出了优异的性能。决策树以目标变量的性质命名，如果目标变量是分类的，称为分类树；如果目标变量是连续的，称为回归树。决策树的目的是根据预测变量开发预测模型。该树是通过根据预测变量之一连续划分数据而形成的。决策树由三种节点组成：根节点、内部节点和叶节点，决策树算法在树的内部节点开发分裂标准[23]。节点的分裂试图最小化节点的杂质。如果分裂在减少杂质方面不能实现任何改进，则该节点不被分裂，并被声明为叶节点。如果分裂能够减少杂质，那么选择提供最大杂质减少的分裂，并且形成两个分支，形成两个新节点。流行的分裂标准是信息增益、基尼指数和增益比。CART 是决策树算法之一，它使用基尼指数来构建二叉树，以选择每个内部节点的分裂变量。则概率分布的基尼系数为：

$$Gini(p)=1-\sum_{k=1}^{K}p_k^2 \tag{6}$$

其中 p_k 表示样本点属于第 k 类的概率，K 表示分类问题有 K 个类。

随机森林是一种集成学习方法。它通过选择给定数据集的子集和随机选择预测变量的子集来生成许多分类树，最后聚集所有模型的结果以获得随机森林。为了得到最终的“多数”分类规则，从自举样本中获得多个分类树。监督机器学习算法将数据分为两部分，即训练数据和测试数据。随机森林和决策树

最重要的特征之一是变量重要性的输出。变量重要性衡量给定变量和分类结果之间的关联程度。随机森林和决策树对可变重要性有四种度量：0 级原始重要性分数、1 级原始重要性分数、准确性下降和基尼指数[24]。

2.5. 支持向量机

支持向量机的原理是从 n 维的数据空间中找到一个超平面将数据集分为两类，分类面的函数形式是 $f(x)=w^{\mathrm{T}}x+b$，其中其中 w^{T},b 为寻找分类面的参数值。假设数据集集有 k 个样本，用 $\{x_i,y_i\},i=1,\cdots,k$ 表示数据集的样本，那么支持向量机寻找超平面的过程可以看作一个凸二次优化问题

$$\min\varphi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}w^{\mathrm{T}}w \tag{7}$$

$$y_i\left(w^{\mathrm{T}}x_i+b\right)-1\geq 0 \tag{8}$$

通过优化的决策函数支持向量机可以寻找到类间距离最大的分离超平面，从而进行分类。

3. 不同分类器下的信贷评估

3.1. 数据处理

本论文使用真实场景的数据，数据来自 Kesci 社区。基于客户的信用卡数据，预测客户是否会守约，从而将客户进行分类：守信的客户和可能会违约的客户。我们将用如下指标来评价客户的信用。由于机器学习模型在训练过程中要求数据必须为数值型，而原始数据集的数据存在缺失值，本文进行数据分析前对数据的缺失值进行了处理，以确保模型可以顺利进行训练，对处理缺失值后的数据进行了汇总，共得到 112410 条数据，其中守信客户有 111912 个，违约客户有 8357 个，其中有 9 个特征值，1 个目标值；将数据集进行拆分，拆分为训练集和测试集，利用伪随机数生成器将数据集打乱，其中 25%的数据集作为测试集。数据集的特征如表 1 所示。

Table 1. Characteristics and categories of credit data sets
表 1. 信贷数据集特征及类别

特征名和目标名	含义	样本均值	变量类别	变量赋值说明
Serious Dlqin 2 yrs	是否逾期	无	分类	0：守信 1：违约
Revolving Utilization Of Unsecured Lines	信用卡和个人信贷额度的总余额	6.08	数值	实际值
Number Of Time 30 - 59 Days Past Due Not Worse	过去 2 年，借款人逾期 30~59 天的次数	0.44	数值	实际值
Debt Ratio	负债比率	352.3	数值	实际值
Monthly Income	月收入	5356	数值	实际值
Number Of Open Credit Lines And Loans	未偿还贷款数量和信贷额度	8.43	数值	实际值
Number Of Times 90 Days Late	借款人逾期 90 天或以上的次数	0.27	数值	实际值
Number Real Estate Loans Or Lines	抵押贷款和房地产贷款的数量	1.01	数值	实际值
Number Of Time 60 - 89 Days Past Due Not Worse	过去 2 年，借款人逾期 60~89 天的次数	0.24	数值	实际值
Number Of Dependents	家庭中的家属人数	0.73	数值	实际值

由于本文中例如年龄、月收入、负债比率的指标数值相差较大，为了消除不同特征下的数据的量纲影响，我们需要对数据进行标准化处理。标准化的处理方式有：最大最小标准化，Z-score 标准化等，我们采用 Z-score 标准化，该标准化转化方法为：

$$x' = \frac{x - \mu}{\sigma} \tag{9}$$

其中 μ 为特征下数据的均值，σ 为特征下数据的标准差，数据处理后呈正态分布，均值为 0，方差为 1。

3.2. 数据特征

根据上文，将信贷人是否违约分为 0 和 1 的两类客户，在借贷人的数据集中，除了此分类属性，其余属性为特征向量，每个借贷人由 (x_i, y_i) 表示，i 为数据集的样本个数，其中 x_i 为 n 维的特征向量，y_i 为 0，1 的二分类集。首先对样本集进行分析，观察样本集内违约客户和守信客户的数量，如图 1 所示，故需对不平衡数据集进行处理，由于样本中守约客户和违约客户的不平衡，使用 BLS 算法对不平衡数据进行处理，以合成信贷问题中的少数类样本。

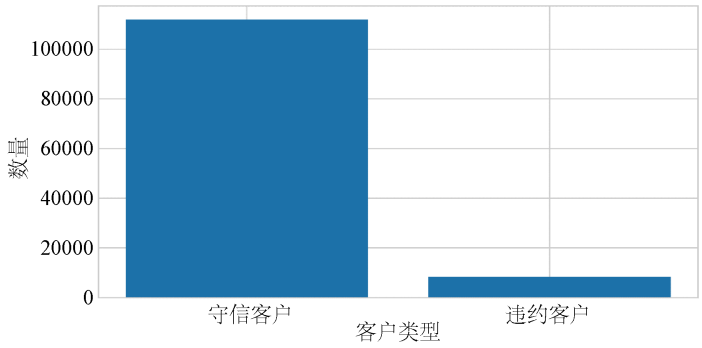

Figure 1. Data set distribution
图 1. 数据集分布情况

对不平衡数据集处理后，构建样本集 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，将样本集进行分割，分割成训练集和测试集。此处的分割比例为 0.25，即样本集中 75%的数据集被随机提取为训练集，剩余 25%数据集作为测试集。

4. 模型分析

4.1. 模型参数设置

本文选取逻辑回归、朴素贝叶斯、决策树、随机森林、支持向量机五种分类器对上述数据进行训练，以观察和评估训练器的性能。由于随机森林和决策树之间存在集合关系，此处对随机森林分类器的参数进行具体说明。随机森林是决策树的集合，它包含了决策树分类器的所有优点，同时还避免了决策树的一些缺点，故随机森林分类器优于决策树分类器。随机森林分类器中有 3 个参数可以调整：n_estimators，max_features，min_sample_leaf，更改这三个参数可以提高模型的预测能力。构造随机森林模型时，保持其他参数默认，分别更改这三个参数，观察分类器的准确性。

首先确定参数 n_estimator，即决策树的个数，更改决策树的个数，观察分类器准确性的变化，如图 2 所示。

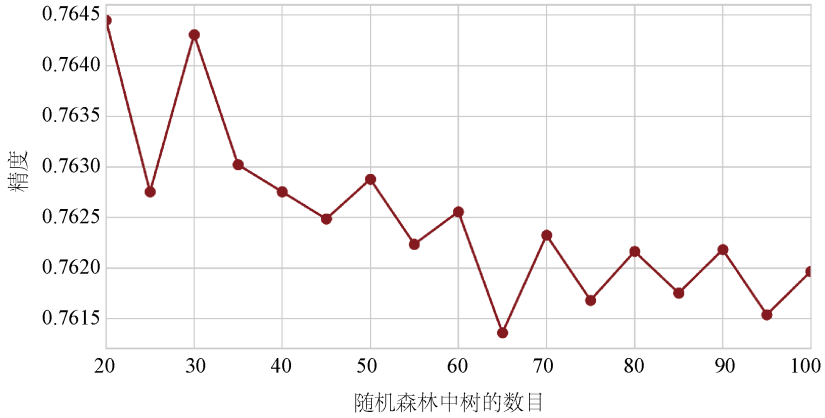

Figure 2. The influence of the precision of random forest simulator on the number of trees
图 2. 随机森林模拟器精度随树的数目的影响

一般来说，树的数量越多，性能越好，但是代码越慢。但我们由图一得出，当随机森林中数目发生改变时，随机森林分类器的精度呈现波动状，波动范围在 0.7615~0.7645 之间，故应设置随机森林分类器的 n_estimators 参数为 20 或 30 为宜。

由于在构造随机森林时，每个决策树对应的数据集都各不相同，决策树的每次划分基于特征值的不同子集，此时我们需要 max_features 的参数对数据集的特征进行选择，max_features 表示随机森林允许在单个树中尝试的最大特征数，当 max_features 的值过大，则随机森林中的决策树相似度非常高，反之，max_features 的值越小，随机森林中的树的差异越大，保持其他值不变，更改参数 max_features 的值，观察精度的变化，如图 3 所示。

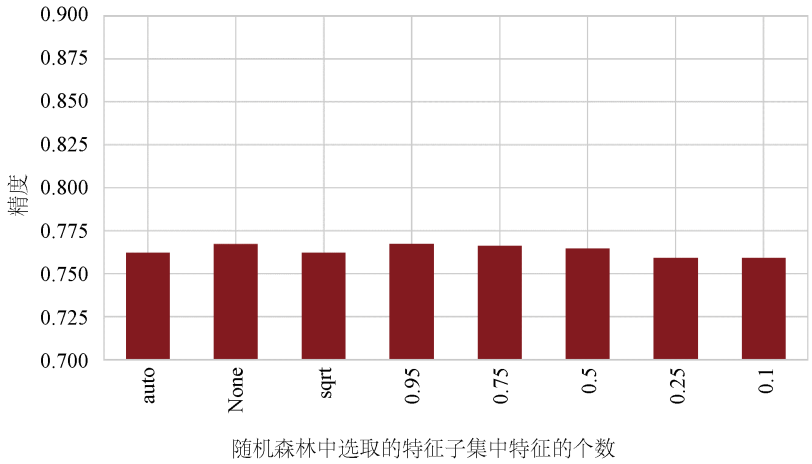

Figure 3. The influence of the precision of random forest simulator on the number of features
图 3. 随机森林模拟器精度随特征个数的影响

由图 2 可以看出，设置随机森林最大特征数为 None 和 0.95 时获得的精度最高，其中 None 表示 max_features 直接等于特征总数，0.95 表示允许随机森林在单独运行中获取 95%的变量。

在构建随机森林模型时，最小样本叶大小十分重要。叶是决策树的末端节点。较小的叶片使模型更容易捕捉列车数据中的噪声，但最小样本叶过小很容易产生过拟合的现象。

如图 4 所示，调整随机森林中最小样本尺寸，当最小样本尺寸小于 25 时，精度随最小样本尺寸的增加明显增大，故在运用随机森林分类器时，应将最小样本尺寸设置为 25~50 之间。

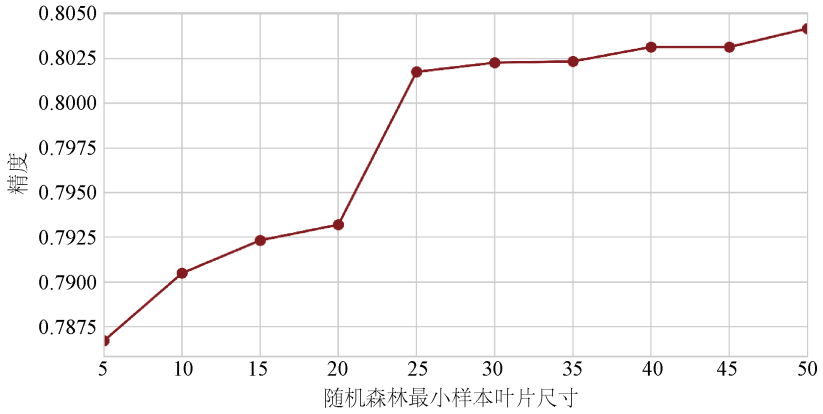

Figure 4. The influence of the precision of random forest simulator on the number of blade size
图 4. 随机森林模拟器精度随叶片尺寸的影响

4.2. 模型性能评估

本文使用的性能评估指标包括精度(accurary)，准确度(precision)，召回率(recall)，f-分数(f-score)，这些评估指标在之前的很多研究中都被使用，每一个指标都有它的优点和局限性，选择指标的组合能够更好的评估分类器的性能。针对本文的二分类问题的评估结果，混淆矩阵是一种全面的表示方法。该矩阵中的每个元素代表对申请人在类别中的位置的正确或不正确预测的数量。

如表 2 所示，TP (True Positive)为真正例，表示客户属性是守信，分类器判断为守信；TN (True Negative)为真反例，表示客户属性是违约，分类器判断为违约；FP (False Positive)为假正例，表示客户属性为守信，分类器判断为违约；FN (False Negative)为假反例，表示客户属性为违约，分类器判断为守信。

Table 2. Description of credit data confusion matrix
表 2. 信贷数据混淆矩阵说明

测试输出	违约客户	守信客户
违约客户	TN	FP
守信客户	FN	TP

精度(accuracy)：精度表示分类器正确分类的样本所占的比例，精度表示模型的成功率，精度可由下式进行定义：

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

准确率(precision)表示真正的正例占被预测为正例样本的比例，在本文中表示被预测为守信中守信客户占总客户的比例，准确度可表示为

$$\text{precision} = \frac{TP}{TP + FP} \tag{11}$$

召回率(recall)反应了分类器避免假反例的性能，本文中召回率度量的守信客户中有多少被预测为守信客户，召回率可表示为

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

将准确率和召回率进行汇总，作调和平均，即得到 f-分数(f-score)，f-score 的计算方式如下式：

$$\text{f} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

ROC 曲线又名受试者工作特征曲线，ROC 曲线考虑了分类器的阈值的变化，呈现了分类器分类的假正例度和真正例度，真正例度即召回率，假正例度 FPR 的计算方式如下式：

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{14}$$

由 ROC 曲线可算出分类器的 AUC 值，ROC 曲线与 FPR 轴线形成的面积即为 AUC 值。AUC 值表示一种概率值，它表示将正样本排在负样本前面的概率，AUC 值越大，表示分类器越有可能将正样本排在负样本前面，意味着分类器能得到更好的分类结果。

将分类器分别对 BLS 算法处理后的数据进行分类，得出分类器性能的比较，如表 3 所示。

Table 3. Evaluation index and score of credit score of different classifications
表 3. 不同分类器信用评分评估指标与评分

分类器	accuracy	precision	recall	f-score
Logistics 回归	0.7797	0.7510	0.8366	0.7916
朴素贝叶斯	0.5272	0.5142	0.9854	0.6758
随机森林	0.8039	0.7922	0.8194	0.8056
支持向量机	0.7678	0.7137	0.8944	0.7939

在信贷数据集中，accuracy 体现了分类器预测正确的概率，precision 表示了分类器对负样本的区分能力，即对违约客户的区分能力，recall 表示分离器对正样本的区分能力，即对守信客户的区分能力，f-score 作为 precision 和 recall 的综合，表现了分类器的稳健，四项评估指标值越高，说明分类器在该项能力上性能越好。从预测结果来看随机森林的分类精度高于逻辑回归分类器和支持向量机分类器，其中朴素贝叶斯分类器的精度表现较差，故朴素贝叶斯分类器不能用于信用评分问题。随机森林在避免假正例方面的性能表现依然最好，模型的准确率高于其他分类器；从客户数据集中找到违约人，在避免假反例的需求下，各分类器的表现良好。其中朴素贝叶斯分类器的性能远强于其他分类器，其次支持向量机分类器强于随机森林和逻辑回归分类器；同时，在综合度量评分下，随机森林的性能也高于支持向量机和逻辑回归分类器，而朴素贝叶斯在综合评分下表现较差。

如图 5~8 的 ROC 图像的结果我们可以看出，随机森林的 ROC 图像更靠近左上方，其次是逻辑回归和支持向量机的 ROC 图像，因此基于决策树的随机森林在此项指标上表现出更好的分类能力，这意味着调节决策树和随机森林的阈值，可以对客户的信用起到很好的分类效果。

从表 4 中的 AUC 值的结果可以看出，随机森林分类器的 AUC 值高于其他分类器的 AUC 值，而朴素贝叶斯的 AUC 值较其他分类器偏低。这表明，通过调节随机森林的阈值，可以得到对数据集更好的分类结果。

Table 4. AUC value of credit score of different classifications
表 4. 不同分类器信用评分 AUC 值

分类器	AUC 值
Logistics 回归	0.8596
朴素贝叶斯	0.7534
随机森林	0.8871
支持向量机	0.8681

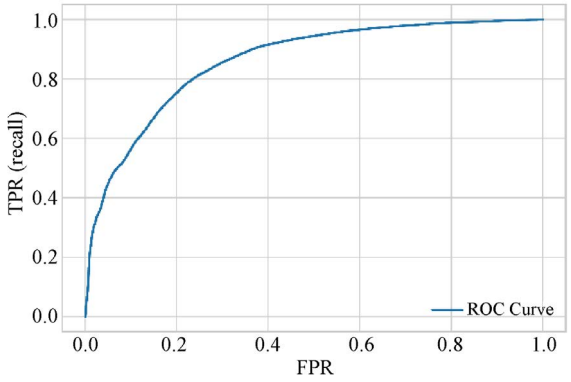

Figure 5. Logistic regression ROC diagram
图 5. 逻辑回归 ROC 图

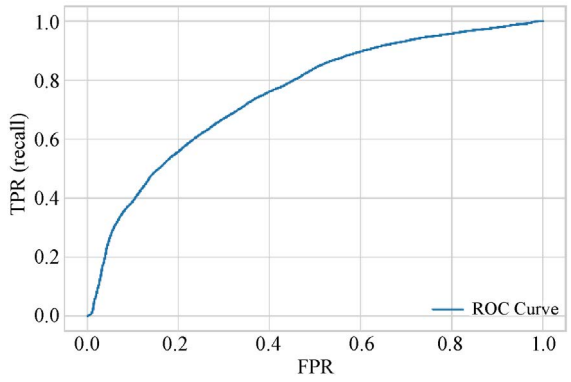

Figure 6. Naïve Bayes ROC diagram
图 6. 朴素贝叶斯 ROC 图

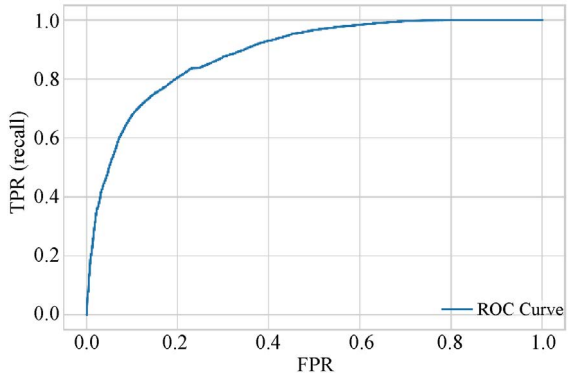

Figure 7. Random forest ROC diagram
图 7. 随机森林 ROC 图

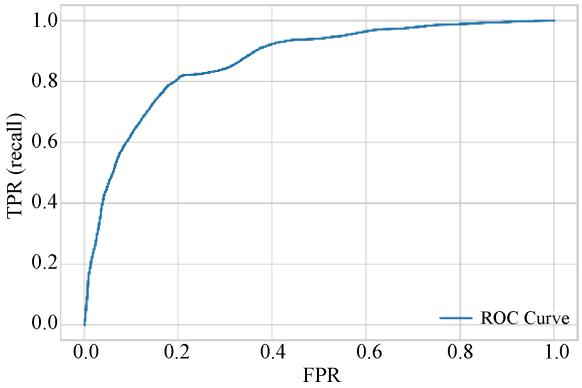

Figure 8. Support machine ROC diagram
图 8. 支持向量机 ROC 图

5. 总结与展望

结果表明，随机森林分类器在保证预测信贷客户是否违约精度的情况下，在区分守信客户和区分违约客户的表现均强于其他分类器，同时随机森林的评估效果较稳健，调节随机森林的阈值可以得到更好的分类效果。由于在信贷客户二分类问题中，及时辨别违约客户能给金融机构带来风险的规避，同时金融机构需要避免误判以保持公信力。因此，随机森林数据集更适合用于信贷客户分类。

机器学习技术作为一项应用潜力巨大的技术，能够被运用到金融风控行业之中，本文将四种分类器：逻辑回归、朴素贝叶斯、决策树、支持向量机应用于金融机构信贷客户的分类问题之中。本文数据表明随机森林分类器对信贷数据集的分类性能较好，支持向量机和逻辑回归分类器对信贷数据集的分类较好，而朴素贝叶斯分类器不适用于对信贷数据集进行分类。未来可将随机森林作为基分类器开发集成模型进一步提高预测准确性和性能。

基金项目

项目名称：信用挖掘为中心的人工智能普惠金融服务技术研究；项目编号：18GLA004。

参考文献

[1] Lawi, A., Aziz, F. and Syarif, S. (2017) Ensemble Gradient Boost for Increasing Classification Accuracy of Credit Scoring. 2017 4*th International Conference on Computer Applications and Information Processing Technology* (*CAIPT*), Kuta Bali, 8-10 August 2017, 1-4. https://doi.org/10.1109/CAIPT.2017.8320700

[2] Imtiaz, S. and Brimicombe, A.J. (2017) A Better Comparison Summary of Credit Scoring Classification. *International Journal of Advanced Computer Science and Applications*, **8**, 1-4. https://doi.org/10.14569/IJACSA.2017.080701

[3] Triki, I. (2016) Credit Scoring Models for a Tunisian Microfinance Institution: Comparison between Artificial Neural Network and Logistic Regression. *Review of Economics & Finance*, **6**, 61-78.

[4] Xia, Y., Zhao, J., He, L., *et al*. (2020) A Novel Tree-Based Dynamic Heterogeneous Ensemble Method for Credit Scoring. *Expert Systems with Applications*, **2020**, Article ID: 113615. https://doi.org/10.1016/j.eswa.2020.113615

[5] Chen, L.-H. and Chiou, T.-W. (1999) A Fuzzy Credit-Rating Approach for Commercial Loans: A Taiwan Case. *Omega*, **27**, 407-419. https://doi.org/10.1016/S0305-0483(98)00051-6

[6] Tsai, C.-F. and Wu, J.-W. (2008) Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring. *Expert Systems with Applications*, **34**, 2639-2649. https://doi.org/10.1016/j.eswa.2007.05.019

[7] Henley, W. and Hand, D.J. (1996) AK-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk. *The Statistician*, **45**, 77-95. https://doi.org/10.2307/2348414

[8] Vapnik, V. (2013) The Nature of Statistical Learning Theory. Springer Science & Business Media, Berlin.

[9] Härdle, W., Moro, R. and Schäfer, D. (2005) Predicting Bankruptcy with Support Vector Machines. In: *Statistical*

Tools for Finance and Insurance, Springer, Berlin, 225-248. https://doi.org/10.1007/3-540-27395-6_10

[10] Pławiak, P., Abdar, M. and Acharya, U.R. (2019) Application of New Deep Genetic Cascade Ensemble of SVM Classifiers to Predict the Australian Credit Scoring. *Applied Soft Computing*, **84**, Article ID: 105740. https://doi.org/10.1016/j.asoc.2019.105740

[11] Maldonado, S., Bravo, C., López, J., *et al*. (2017) Integrated Framework for Profit-Based Feature Selection and SVM Classification in Credit Scoring. *Decision Support Systems*, **104**, 113-121. https://doi.org/10.1016/j.dss.2017.10.007

[12] Kiran, S., Kumar, N., Guru, J., *et al*. (2018) Credit Card Fraud Detection Using Naïve Bayes Model Based and KNN Classifier. *International Journal of Advance Research*, *Ideas and Innovations in Technology*, **4**, 44-47.

[13] Okesola, O.J., Okokpujie, K.O., Adewale, A.A., *et al*. (2017) An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach. 2017 *International Conference on Computational Science and Computational Intelligence* (*CSCI*), Las Vegas, 14-16 December 2017, 228-233. https://doi.org/10.1109/CSCI.2017.36

[14] Chang, Y.-C., Chang, K.-H., Chu, H.-H., *et al*. (2016) Establishing Decision Tree-Based Short-Term Default Credit Risk Assessment Models. *Communications in Statistics—Theory and Methods*, **45**, 6803-6815. https://doi.org/10.1080/03610926.2014.968730

[15] Chawla, N.V., Bowyer, K.W., Hall, L.O., *et al*. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[16] Han, H., Wang, W.-Y. and Mao, B.-H. (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *International Conference on Intelligent Computing*, Hefei, 23-26 August 2005, 878-887. https://doi.org/10.1007/11538059_91

[17] Wang, K.J., Adrian, A.M., Chen, K.H., *et al*. (2015) A Hybrid Classifier Combining Borderline-SMOTE with AIRS Algorithm for Estimating Brain Metastasis from Lung Cancer: A Case Study in Taiwan. *Computer Methods and Programs in Biomedicine*, **119**, 63-76. https://doi.org/10.1016/j.cmpb.2015.03.003

[18] Zhang, J. and Li, X. (2017) Phishing Detection Method Based on Borderline-Smote Deep Belief Network. *International Conference on Security*, *Privacy and Anonymity in Computation*, *Communication and Storage*, Guangzhou, 12-15 December 2017, 45-53. https://doi.org/10.1007/978-3-319-72395-2_5

[19] Berkson, J. (1944) Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, **39**, 357-365. https://doi.org/10.1080/01621459.1944.10500699

[20] Luo, C. (2019) A Comprehensive Decision Support Approach for Credit Scoring. *Industrial Management & Data Systems*, **120**, 280-290. https://doi.org/10.1108/IMDS-03-2019-0182

[21] Akkoç, S. (2019) Exploring the Nature of Credit Scoring: A Neuro Fuzzy Approach. *Fuzzy Economic Review*, **24**, 3-24. https://doi.org/10.25102/fer.2019.01.01

[22] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[23] Tayefi, M., Tajfard, M., Saffar, S., *et al*. (2017) hs-CRP Is Strongly Associated with Coronary Heart Disease (CHD): A Data Mining Approach Using Decision Tree Algorithm. *Computer Methods and Programs in Biomedicine*, **141**, 105-109. https://doi.org/10.1016/j.cmpb.2017.02.001

[24] Lu, Z.Q.J. (2010) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the Royal Statistical Society Series A*, **173**, 693-694. https://doi.org/10.1111/j.1467-985X.2010.00646_6.x