Hans Journal of Data Mining 数据挖掘, 2021, 11(2), 50-58
Published Online April 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.112006

Hans汉斯

# 改进SMOTE算法在Logistic回归信用评分模型中的应用

**许芷慧，杨立洪**

华南理工大学数学学院，广东 广州
Email: 2467205564@qq.com

收稿日期：2021年2月20日；录用日期：2021年3月23日；发布日期：2021年3月30日

## 摘 要

**信用评分模型是商业银行贷前审批的重要应用模型，它通过提前识别出高风险客户来降低银行遭受信贷违约和欺诈的风险。Logistic回归模型作为最广泛使用的信用评分模型，对于信贷数据样本不平衡的特点较为敏感，若不改善样本不平衡问题，将会使模型的分类性能欠佳。为此，本文结合Logistic回归原理，提出了考虑变量重要性来合成辅助样本的改进SMOTE过采样算法(FW_SMOTE)，通过与传统SMOTE、一些经典的改进SMOTE算法，如Borderline-SMOTE和ADASYN做实验对比，发现FW_SMOTE过采样算法使Logistic回归信用评分模型的效果有所改善，具有一定的应用价值。**

## 关键词

**SMOTE算法，过采样，变量权重，Logistic回归**

# Application of Improved SMOTE Algorithm in Logistic Regression Credit Scoring Model

**Zhihui Xu, Lihong Yang**

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: 2467205564@qq.com

Received: Feb. 20th, 2021; accepted: Mar. 23rd, 2021; published: Mar. 30th, 2021

## Abstract

**Credit scoring model is an important application model for pre-loan approval of commercial banks. It can help the bank reduce the risk of credit default and fraud by identifying high-risk cus-**

文章引用: 许芷慧, 杨立洪. 改进 SMOTE 算法在 Logistic 回归信用评分模型中的应用[J]. 数据挖掘, 2021, 11(2): 50-58.
DOI: 10.12677/hjdm.2021.112006

**tomers in advance. Logistic regression model, as the most widely used credit scoring model, is sensitive to the imbalance of credit data samples. If the problem of samples imbalance is not improved, the classification performance of the model will be poor. To this end, combined with Logistic regression principle, we propose an improved SMOTE algorithm which produces the auxiliary sample through the method of feature weighting synthesis (FW_SMOTE), and compare it with traditional SMOTE, some classic improved SMOTE algorithm, such as Borderline-SMOTE and ADASYN by experiment contrast, finding that FW_SMOTE makes the Logistic regression performance of credit scoring model improve and has a certain application value.**

## Keywords

**SMOTE Algorithm, Sampling, Feature Weighting, Logistic Regression**



## 1. 引言

信用卡是当今非现金信贷消费支付的重要手段之一。由于信用风险和欺诈风险的存在，信用卡违约支付行为屡见不鲜，为商业银行的经营带来不利影响。因此，在信用审批阶段，构建有效的信用评分模型来对高风险客户进行提前识别至关重要。

目前对于信用评分模型的构建已经具有一定的研究基础。随着人工智能时代的到来，决策树、Logistic回归、xgboost 等机器学习模型在金融风控中逐渐广泛使用。而 Logistic 回归模型由于可解释性高、泛化性能较好，是信用评分较广泛使用的模型。但由于信用评分数据通常具有高风险样本较少的样本不平衡特点，而 Logistic 回归模型的损失函数是考虑了所有训练样本分类结果的似然函数，若不做处理将会使得模型对高风险样本错判率提高，因此需要寻求有效的适合 Logistic 回归信用评分模型的不平衡样本处理方法。

当前对于不平衡样本的处理，主要可以分为两种，一种是数据层面的欠采样、过采样算法，另一种是分类算法层面的代价敏感学习方法、一分类、集成算法等。分类算法层面的不平衡样本一般难度较大，相比起数据层面的处理复杂度更高。在数据层面的处理算法中，欠采样是通过筛选多数类样本来使总体样本达到平衡，但会损失多数类样本的部分信息，导致多数类样本的分类准确率下降。而过采样是通过扩增少数类样本来改善样本不平衡，可以保留样本更多信息，同时增加样本多样性[1]。而 SMOTE 是比较经典的一种过采样方法。它通过在原始少数类样本与其一个随机同类近邻样本之间进行插值来产生新样本[2]，但同时也存在产生噪声样本的问题。传统 SMOTE 算法虽然相比简单随机过采样更能增加样本的多样性，但由于无论是根样本或辅助样本的选取均是一定条件下简单随机抽取的，因此 SMOTE 算法衍生出的样本的多样性其实具有较大的随机性。若抽取的根样本或其辅助样本代表性不足，或处于两类样本的边界，则衍生出来的样本的噪音特性可能多于有效特性，反而会为分类器的学习带来干扰[3]。

在 SMOTE 改进的研究中，关于根样本与辅助样本的选取，前人提出过不少经典的优化算法。Han等[4]认为处于分类边界的样本对于分类具有更大的重要性，提出只对位于边界但非噪声的少数类样本进行 SMOTE 衍生的 Borderline-SMOTE 算法。HE Haibo 等[5]在 2008 年提出 ADASYN 算法，根据原始少数类样本的分布特点，认为 k 近邻中异类样本占比越大的少数类样本是越难学习的样本，以其作为根样

本时也需要合成越多的新样本。ZHU Tuanfai 等[6]在 2017 年提出了 SMOM 算法，根据过泛化风险的大小，赋予根样本的 k 近邻同类样本不同的选择权重，权重越大代表该样本被选为辅助样本的概率越大，从而改善过泛化问题。Li 等[7]在 2015 年提出结合 Lasso 方法筛选重要变量集合，为集合中的变量赋予权重值为 2，其他变量权重值为 1，然后对进行变量赋权后的不平衡样本集进行传统 SMOTE 过采样。

之前的学者关于 SMOTE 算法改进主要存在两个问题，一个是多数基于对样本分类的普适性认识进行改进，而较少从某个分类器的原理特点出发进行针对性优化；另一个是在关于辅助样本选取的改进上，虽然有赋予样本变量不同权重的尝试[7]，但基本没有从根据变量权重将多个近邻样本加权合成一个辅助样本的角度切入进行尝试和改进。

为此，本文以 Kaggle 德国信用卡风险样本为实验数据，结合 Logistic 回归模型的原理特点，提出一种考虑变量重要性来合成辅助样本的 SMOTE 过采样方法，针对性改善 Logistic 回归信用评分模型构建过程中受到样本不平衡问题的影响，最后基于实验数据与传统的 SMOTE、Borderline-SMOTE、ADASYN 在 Logistic 回归信用评分模型上进行对比分析。

## 2. 相关算法

### 2.1. Logistic 回归分类算法

Logistic 回归模型[8]是对数线性模型，常用于解决二分类或多分类问题，Logistic 回归的函数形式为

$$P(y=1|x)=\text{sigmod}\left(\beta_0+\beta_1^{\mathrm{T}}x\right)=\frac{\mathrm{e}^{\beta_0+\beta_1^{\mathrm{T}}x}}{1+\mathrm{e}^{\beta_0+\beta_1^{\mathrm{T}}x}}$$

Sigmod 映射函数不改变自变量与函数值的线性相关的正负方向，因此 Logistic 回归的系数能直观体现自变量对输出概率值的影响程度。

Logistic 回归函数的参数估计方法采用极大似然估计法来进行参数估计，估计步骤如下：

Step1：将 $Y$ 的概率函数改写为

$$P(Y|x)=P^{Y}(1-P)^{1-Y},\ Y=0,1$$

Step2：得到似然函数

$$L=\prod P(y_i|x_i)=\prod P^{y_i}(1-P)^{1-y_i}$$

Step3：对 $L$ 取自然对数并对 $\beta_0$ 和 $\beta$ 求偏导，使偏导为 0，求解出极大似然估计量 $\hat{\beta}_0$ 和 $\hat{\beta}$。

### 2.2. SMOTE 过采样算法

SMOTE 算法是 Chawla 等[2]在 2002 年提出的一种对不平衡数据进行预处理的方法，区别于随机过采样，SMOTE 是基于少数类样本的 $k$ 近邻同类样本的线性插值过采样方法，而并非简单复制原始少数类样本，一定程度上使样本的类分布更均衡，减少分类器过拟合的风险。具体的 SMOTE 过采样算法如下：

输入：少数类样本集 $X_1$，采样倍数 $N$，近邻个数 $k$。

输出：合成的新少数类样本 $X_{new}$ 集。

Step1：从 $X_1$ 中随机抽取少数类样本 $\{x_1,x_2,\cdots,x_T\}$。

Step2：对于 $\{x_1,x_2,\cdots,x_T\}$ 中每一个样本 $x_i$，$i=1,2,\cdots,T$，计算 $x_i$ 在 $X_1-\{x_i\}$ 的 k 近邻样本，并将其放进集合 $X_{ik}$ 中。

Step3：根据采样倍率 $N$，从 $X_{ik}$ 中随机抽取若干近邻样本 $x_{ij}$，取[0, 1]之间的随机数 $r$，令新样本为

$$x_{new} = x_i + r \times \left( x_{ij} - x_i \right)$$

Step4：将 $x_{new}$ 加入 $X_{new}$。

## 3. 基于改进 SMOTE 算法的 Logistic 回归信用评分模型

### 3.1. 分箱与 WOE 编码

基于 SMOTE 的过采样方法无法直接对字符串、二值等属性的离散型变量直接进行插值操作，一般需要先对离散型变量进行编码。Logistic 回归信用评分模型在特征处理上，为增强模型鲁棒性，通常会对变量进行分箱和 WOE 编码[9]。

#### 3.1.1. 基于高风险样本占比差异最大化的分箱

输入：待分箱数据集 $X$，离散型变量 $col$，最大分箱数 $n$，每箱最少样本数占总体样本比例 $p$。

输出：分箱结果。

初始化：当前箱数 $n_{bin}$ 为变量 $col$ 的取值个数。

Step1：若 $n_{bin} > n/2$，计算变量$col$各个箱的高风险样本占比并按降序排序，并将差值最小的两箱进行合并，$n_{bin} = n_{bin} - 1$；否则，直接结束分箱。

Step2：若 $n_{bin} > n$，回到 step1；否则，到 step3。

Step3：若当前存在箱的最小样本占总体样本比例小于 $p$ 且当前箱数大于 $n/2$，将该箱与其相邻高风险样本占比差值最小的箱合并，直至所有箱的最小样本占总体样本比例不小于 $p$ 或当前箱数不大于 $n/2$，然后结束分箱；否则，直接结束分箱。

#### 3.1.2. WOE 值计算

分箱后的 WOE 值计算公式如下：

$$\mathrm{WOE}_i = \ln\left( \frac{pyi}{pni} \right)$$

其中，$pyi$ 是分箱中高风险样本占总体高风险样本的比例，$pni$ 是箱中低风险样本占总体低风险样本的比例。

### 3.2. 数据标准化与度量

数据标准化方法：离差标准化

$$x_i^* = \frac{x_i - \min\limits_{1 \le j \le n} x_j}{\max\limits_{1 \le j \le n} x_j - \min\limits_{1 \le j \le n} x_j}, i = 1, 2, \cdots, n$$

样本距离度量：欧氏距离

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

### 3.3. 预训练：计算 Logistic 回归的变量权重

设训练样本集为 $X_{train}$，训练样本集的标签为 $y_{train}$，样本的变量为 $F_l = (f_1, \cdots, f_n)$。

输入：训练集 $(X_{train}, y_{train})$。

输出：变量权重向量 $W = (w_1, \cdots, w_n)$。

Step1：以 $F_l$ 为输入变量，训练 Logistic 回归模型并调节参数。

Step2：若 Logistic 回归模型 $k$ 折交叉验证准确率大于等于 60%，提取输入变量回归系数 $A=(a_1,a_2,\cdots,a_n)$；否则，回到 Step1。

Step3：通过归一化 $A$ 获取变量权重向量 $W=(w_1,\cdots,w_n)$。

### 3.4. 考虑变量重要性的 SMOTE 过采样(FW_SMOTE)——基于 Logistic 回归分类器

不同的分类器具有不同的分类原理，因此改进 SMOTE 过采样算法对于不同分类器也可能会有不同的适用性。Logistic 回归模型是对数线性模型，回归系数可以直观反映变量对输出概率值的相对影响程度，一般对分类越重要的变量变动越大，其对逻辑回归输出结果的直接影响也越大。此外，由于损失函数是基于所有训练样本的似然函数，迭代求解回归系数的过程受多数类样本的错分影响更大，所以样本不平衡会对参数估计造成较大影响。因此，考虑到 Logistic 回归的特点和原始 SMOTE 算法容易合成噪声的问题，本文提出一种基于变量重要性来加权合成辅助样本的 FW_SMOTE 算法。

FW_SMOTE 算法的基本思想就是在辅助样本的选择时，首先通过预训练获取变量权重，然后以整体变量为维度选取 $k$ 近邻样本，在 $k$ 近邻样本中依次以各个变量为单独维度选取各个变量维度上与根样本最相近的样本，再按照各个变量的权重将各变量维度上的最相似样本加权合成辅助样本，最后将根样本和辅助样本进行插值得出新样本。FW_SMOTE 算法步骤如下：

输入：少数类样本集合 $X_1$，需要扩增的样本数量 $num$，参考的近邻样本数量 $k$。

输出：扩增的样本 $X_{new}$。

初始化计数变量 $count=0$。

Step1：预训练获取变量权重 $W=(w_1,\cdots,w_n)$。

Step2：从 $X_1$ 中随机选取一个少数类样本 $x_i$，令 $X_c=X_1-\{x_i\}$。

Step3：计算 $x_i$ 与 $X_c$ 中所有候选样本的整体距离，将距离最小的前 $k$ 个近邻样本加入集合 $X_i$ 中。

Step4：将变量按权重进行降序排序，依次取权重最大(设第 $p$ 大的权重记作 $w^{(p)}$)的变量 $f^{(p)}$，以 $f^{(p)}$ 为单一维度来计算 $x_i$ 与 $X_i$ 中样本的距离，取距离最接近的样本 $x_i^{(p)}$ 作为该变量维度下的最优插值样本。以此类推，直到选完权重最小的变量对应的最优插值样本，可得到 $X_w=\left(x_i^{(1)},\cdots,x_i^{(n)}\right)$。

Step5：进行插值。选取[0，1]之间的随机数 $r$，令 $x_{new}=x_i+r*\sum_{k=1}^{n}w_k*\left(x_i^{(k)}-x_i\right)$，将 $x_{new}$ 加入 $X_{new}$。

Step6：$count=count+1$，若 $count\geq num$，结束插值；否则，返回 step2。

### 3.5. FW_SMOTE_Logistic 回归信用评分模型

#### 3.5.1. 建模流程

1) 获取信用评分数据。

2) 数据预处理：包括数据清洗、3.2 提到的高风险样本比例差异最大化为原则的分箱、WOE 编码处理、数据标准化处理。

3) 划分训练/测试集：为进行五折交叉验证，将数据集划分为五份，每次取四份为训练集，一份为测试集。

4) 少数类样本衍生:采用 3.4 提出的 FW_SMOTE 过采样方法对训练集的少数类样本进行扩增。

5) 训练 Logistic 回归分类模型：将扩增样本后的训练集数据放进 Logistic 回归模型中进行训练，调节 Logistic 回归模型的参数。

6) 模型评价：将测试集数据输入训练后的 Logistic 回归模型，通过 3.5.2 模型评价方案考察模型分类效果。

7) 建立评分映射：对输出概率进行评分映射。(由于本文主要是研究改进 SMOTE 算法对 Logistic 回归信用评分模型中的分类性能的改善效果，因此实验中的将省略评分映射的步骤。)

### 3.5.2. 模型评价方案

为检验本文提出的 FW_SMOTE 过采样方法对 Logistic 回归信用评分模型中的分类性能的改善效果，将同时对比不做重采样、SMOTE 过采样、BorderlineSMOTE 过采样、ADASYN 过采样方法的 Logistic 回归信用评分模型中的分类效果。模型评价的从以下三方面进行：

1) 模型分类准确性评价

混淆矩阵：如表 1。

**Table 1.** Confusion matrix
**表 1.** 混淆矩阵

| | 预测为负 | 预测为正 |
|---|---|---|
| 实际为负 | TN (真负类数量) | FP (假正类数量) |
| 实际为正 | FN (假负类数量) | TP (真正类数量) |

准确率：可以用来衡量整体的预测精度，即被判别正确的样本数量占整体样本数量的比值。

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

召回率：可以评价原始样本中的高风险(正)样本有多少被分类器鉴别出来。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2) 模型泛化能力评价

通过比较训练集/测试集的精度差异(即训练集/测试集在模型评价指标表现的差异)，若两者相差不足 5%，可以认为模型过拟合风险较低，具有较好的泛化性能。

3) 模型综合性能评价

$F_{\text{score}}$：采取 F_score 作为综合评价指标，其中 $\beta$ 值代表了查准率和召回率的相对重要性，设置 $\beta$ 越大，认为召回率越重要。本文将采取 $F_1$ 和 $F_2$ 作为综合评价指标。

$$F_\beta = \left(1 + \beta^2\right) \times \frac{\text{Precision} \times \text{Recall}}{\left(\beta^2 \times \text{Precision}\right) + \text{Recall}}$$

## 4. 实验过程与结果分析

本实验将对 Logistic 回归信用评分模型的分类效果进行验证与对比分析。

### 4.1. 实验数据处理

实验环境为用 Anaconda 搭建的 Python 编译环境，使用编译器为 Spyder，使用编译语言为 python3.6。实验数据来自 Kaggle 平台公开的德国信用卡风险数据集，含有高风险样本 300 个，低风险样本 700 个，包括类别标签一共 10 个变量，数据属性与缺失情况如表 2。

为保证模型精度，对连续数值型变量不做分箱，字符串型变量中，除类别标签 Risk 以外，均以负样本占比最大化为原则做分箱和 WOE 编码，类别标签转换为 0~1 变量，高风险样本标签值为 1，低风险样本标签值为 0。对于缺失值，将在分箱时单独作为一种取值处理。

**Table 2.** Variable attributes and missing situation
**表 2.** 变量属性与缺失情况

| 变量名 | 变量属性 | 缺失情况 |
|---|---|---|
| Age | Int64 | 0 |
| Sex | Object | 0 |
| Job | Int64 | 0 |
| Housing | Object | 0 |
| Saving accounts | Object | 183 |
| Checking account | Object | 394 |
| Credit amount | Int64 | 0 |
| Duration | Int64 | 0 |
| Purpose | Object | 0 |
| Risk (类别标签) | Object | 0 |

## 4.2. 归一化后的变量权重

由图 1 可以发现，9 个变量在 Logistic 回归模型的重要性有较大的差异，其中变量 Duration 的重要性最高，因此在 FW_SMOTE 的过采样中，在 Duration 维度上与根样本越接近的近邻样本将会分得越大的权重，而最后合成的新样本在重要性越大的变量维度的表现将与原始少数类样本越相似，相比传统 SMOTE 更能反映原始少数类样本的在重要变量上的取值特点。

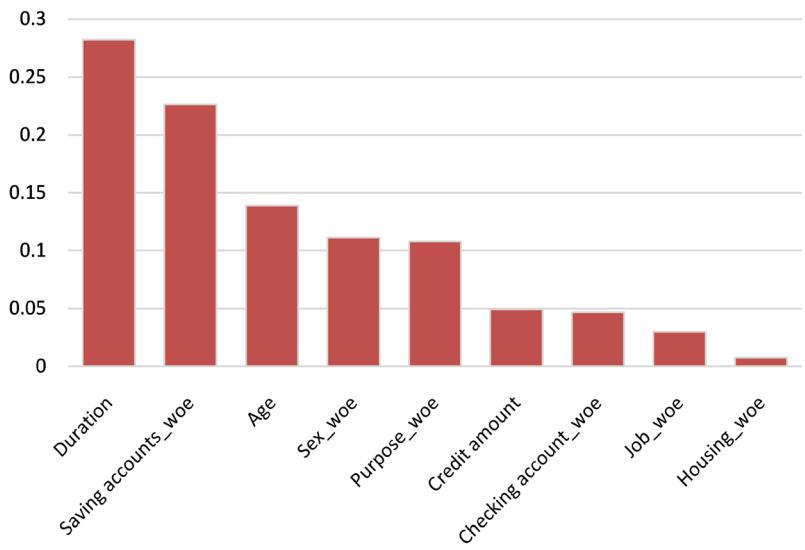


**Figure 1.** A demonstration of the weight of variables obtained from a pre-training
**图 1.** 一次预训练得出的变量权重展示

## 4.3. 实验结果

为减少结果受数据集划分差异的影响，本文实验采用五折交叉验证，表 3 是取定近邻数$k$为 6 的五折交叉验证的结果。

**Table 3.** Five fold cross test results
**表 3.** 五折交叉检验结果

| | Logistic | | SMOTE_Logistic | | FW_SMOTE_Logistic | | Borderline-SMOTE_Logistic | | ADASYN_Logistic | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 训练集 | 测试集 | 训练集 | 测试集 | **训练集** | **测试集** | 训练集 | 测试集 | 训练集 | 测试集 |
| 准确率 | 74.47% | 73.00% | 73.12% | 68.30% | **72.50%** | **68.90%** | 72.05% | 68.40% | 70.17% | 67.50% |
| 高风险样本召回率 | 40.55% | 40.97% | 77.11% | 72.42% | **77.01%** | **75.45%** | 77.07% | 74.81% | 73.72% | 75.00% |
| F_score ($\beta=1$) | | 0.4763 | | 0.5779 | | **0.5927** | | 0.5873 | | 0.5807 |
| F_score ($\beta=2$) | | 0.4340 | | 0.6576 | | **0.6802** | | 0.6743 | | 0.6717 |

从表 3 中的测试集表现可以看出，本文提出的 FW_SMOTE_Logistic 模型在高风险样本召回率和 F_score 上比不做过采样的 Logistic 回归模型有明显提高，有效改善了 Logistic 回归对于样本不平衡敏感的问题，对比基于经典 SMOTE、BorderlineSMOTE 和 ADASYN 的 Logistic 回归模型，FW_SMOTE_Logistic 效果也更好一点，而且训练集和测试集的表现相差不超过 5%，过拟合风险较小。为进一步验证本文 FW_SMOTE_Logistic 模型在实验数据上的适用性和稳定性，将进行稳健性检验。

## 4.4. 稳健性检验

通过修改 FW_SMOTE、经典 SMOTE、BorderlineSMOTE 和 ADASYN 算法的近邻数 $k$，观察五折交叉验证下的召回率、$F_2$ 值评价指标表现：

由稳健性检验结果图 2 和图 3 可以看出，基于当前的实验数据，FW_SMOTE 对 Logistic 回归信用评分模型的改善比其余三种经典改进 SMOTE 算法略优，说明其过采样性能在 Logistic 回归信用评分模型应用上与传统的 SMOTE、Borderline-SMOTE 和 ADASYN 算法有较高的可比性，在 Logistic 回归信用评分模型改善样本不平衡问题上具有一定应用价值。

由于本文只选取了德国信用卡数据集作为实验数据，FW_SMOTE 算法表现评估还是有一定的数据局限性。但对不同数据集的模型构建本来就没有固定最优的模型或算法一说，针对不同的数据集应尝试几

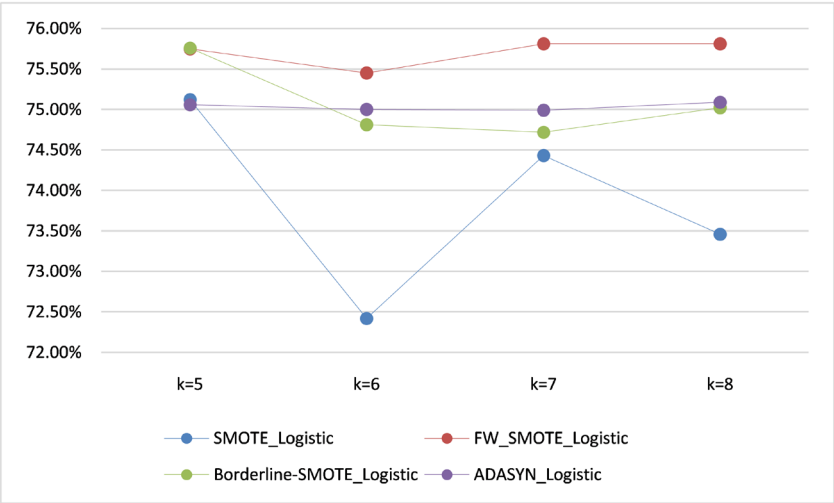


**Figure 2.** High risk sample recall rate in test set with different nearest neighbor numbers
**图 2.** 不同近邻数下测试集高风险样本召回率

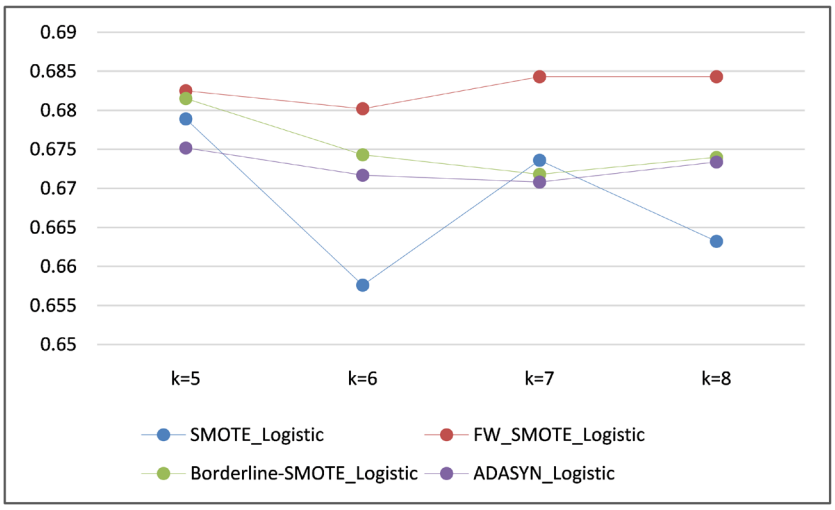


**Figure 3.** Test set $F_2$ score under different nearest neighbor numbers
**图 3.** 不同近邻数下测试集 $F_2$ 值

种算法择优使用。本文提出的 FW_SMOTE 算法，旨在为 Logistic 回归信用评分模型在解决样本不平衡问题提供多一种有一定合理性和参考价值的过采样方法。

## 5. 结论

Logistic 回归信用评分模型是构建信用评分模型广泛使用的模型，但由于信用风险数据集一般具有高风险样本较少的特点，若不改善样本不平衡问题，将会导致 Logistic 回归模型对于高风险样本的分类性能欠佳。而本文提出的 FW_SMOTE 过采样方法，结合 Logistic 回归的原理特点，在辅助样本选择上考虑了变量重要性来进行加权合成。基于本文实验数据对比发现，FW_SMOTE 在 Logistic 回归信用评分模型上的应用性能比 SMOTE、Borderline-SMOTE、ADASYN 有所提升，具有一定的应用价值，同时实验结果也存在一定数据局限性。

## 参考文献

[1] 向鸿鑫, 杨云. 不平衡数据挖掘方法综述[J]. 计算机工程与应用, 2019, 55(4): 1-16.

[2] 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

[3] 石洪波, 陈雨文, 陈鑫. SMOTE 过采样及其改进算法研究综述[J]. 智能系统学报, 2019, 14(6): 1073-1083.

[4] Han, H., Wang, W.-Y. and Mao, B.-H. (2005) *Borderline*-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11538059_91

[5] He, H., Bai, Y., Garcia, E.A., *et al.* (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks*, 1322-1328.

[6] Zhu, T., Lin, Y. and Liu, Y. (2017) Synthetic Minority Oversampling Technique for Multiclass Imbalance Problems. *Pattern Recognition*, **72**, 327-340. https://doi.org/10.1016/j.patcog.2017.07.024

[7] Li, X., Zou, B., Wang, L., Zeng, M., Yue, K., Wei, F., *et al.* (2015) A Novel LASSO-Based Feature Weighting Selection Method for Microarraydata Classification. *Proceedings of* 2015 *IET International Conference on Biomedical Image and Signal Processing*, Beijing, 1-5.

[8] 廖芹, 郝志峰, 陈志宏. 数据挖掘与数学建模[M]. 北京: 国防工业出版社, 2010: 24-28.

[9] 梅子行. 智能风控[M]. 北京: 机械工业出版社, 2020: 28-33.