Computer Science and Application 计算机科学与应用, 2020, 10(11), 2049-2057
Published Online November 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1011216

Hans汉斯

基于逻辑回归的不平衡数据算法适用性研究

李超杰[1]，温　磊[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, wl13365771418@163.com

收稿日期：2020年11月3日；录用日期：2020年11月18日；发布日期：2020年11月25日

摘　要

逻辑回归模型容易受到不平衡数据的影响，本文主要探究了随机欠采样法、Border Line-Smote (BLS)过采样法、自适应综合过采样法(Synthetic Minority Oversampling Technique)等三种不平衡数据算法对逻辑回归模型的适用情况。利用逻辑回归模型分别对三种方法平衡之后的数据，处理之后发现BLS过采样法得出的各项指标最优，ADASYN过采样法得出的各项指标最差，最终得出BLS过采样法更适用于逻辑回归模型的不平衡数据集的处理。

关键词

逻辑回归，随机欠采样法，BSL过采样法，ADASYN过采样法

Research on the Applicability of Unbalanced Data Algorithm Based on Logistic Regression

Chaojie Li[1], Lei Wen[2]

[1]Bank of Communications Jiangsu Branch, Nanjing Jiangsu
[2]Southeast University, Nanjing Jiangsu
Email: lichaojie@bankcomm.com, wl13365771418@163.com

Received: Nov. 3rd, 2020; accepted: Nov. 18th, 2020; published: Nov. 25th, 2020

Abstract

The logistic regression model is susceptible to the impact of unbalanced data. This paper mainly

文章引用: 李超杰, 温磊. 基于逻辑回归的不平衡数据算法适用性研[J]. 计算机科学与应用, 2020, 10(11): 2049-2057.
DOI: 10.12677/csa.2020.1011216

explores the applicability of three kinds of unbalanced data algorithms, including stochastic under-sampling, Border Line-Smote oversampling (BLS) method, and Synthetic Minority Oversampling Technique, to the logistic regression model. By using logistic regression model to process the balanced data of the three methods, it was found that the indicators obtained by BLS oversampling method were the best and the indicators obtained by ADASYN over-sampling method were the worst. Finally, it was concluded that BLS oversampling method was more suitable for the processing of unbalanced data sets of logistic regression model.

Keywords

Logistic Regression, Random Over-Sampling, Border Line-Smote Method, ADASYN Method

1. 引言

将客户进行准确的信用分类对金融的发展极为重要，信贷业务的数据集往往具有不平衡性，而数据集的不平衡会造成信用评级的准确率虚高给金融机构带来损失[1]，不利于金融的发展。所谓的不平衡数据集指的是在一个数据集中，某些类型样本数量极多，某些类型样本极少[2]，当分类器处理不平衡数据集时虽然会获得较高的准确率，但是容易把类型少的样本错分为类型多的样本。解决不平衡数据集带来的问题是金融机构对客户进行信用风险预测之前必不可少的步骤。

随着各领域学者对不平衡数据集的不断研究，不平衡数据集的处理方法也越来越丰富，如今主要分为三类方法：基于分类算法、基于数据采样法、分类算法－数据采样综合法[3]。

顾东晓等提出的重采样法可以分为欠采样和过采样两个大类，是处理不平衡数据集的有效方法之一[4]；Chawle 等人于 2005 年提出了 SMOTE 过采样技术，SMOTE 技术可以在很大程度上降低过拟合发生的概率[5]；Madireddi Vasu 和 Vadlamani Ravi 于 2011 年研究表明利用 K-反向近邻法和 K-均值聚类法相结合可以降低数据集中多数类样本的影响[6]；李辉等人在 2011 年提出了一种新的上采样方法和预测方法来校正不平衡样本[7]；G. Ganesh Sundarkumar 等人在 2015 年时提出了利用 K-反向近邻法和支持向量机相结合的方法来纠正数据不平衡所产生的问题[8]；Jerzy Blaszczyński 等人在 2015 年证明了用一些集成方法处理不平衡数据集比普通方法更有效[9]；Jingjun Bi 等人 2018 年创造出了多样化纠错输出码方法可以用来有效解决类不平衡[10]；Anahita Namvar 等人 2018 年研究表明随机森林和随机欠采样法相结合可以有效解决不平衡数据集造成的偏差[11]；高阳、刘其成和牟春晓三人在《基于蚁群聚类的不平衡数据过采样方法》一文中表明 ACC-SMOTE 采样算法可以明显提高不平衡数据集的分类精度[12]；蒋华和江日晨等人 2020 年用 ADASYN 和 SMOTE 相结合成功处理了不平衡数据集[13]。

各领域的学者大部分研究的是处理不平衡数据集的方法[14]，但很少关注何种处理数据集的方法适合分类器。逻辑回归模型作为金融机构常用的信用风险预测模型很容易受到不平衡数据集的影响，通过查找文献得知前人并没有研究何种处理不平衡数据的方法更适合逻辑回归模型。随机欠采样法、Border Line-Smote 过采样法、自适应综合过采样法(ADASYN)等三种方法是比较典型也是比较热门的对不平衡

数据集进行平衡的方法，故此本文以这三种方法为例探究哪种方法更适用于逻辑回归模型。

2. 不平衡数据集及其处理方法介绍

不平衡数据集指的是在一个数据集中某些样本数量非常多(简称为多数类样本)，某些样本数量非常少(简称为少数类样本)的现象。不平衡数据导致的问题主要是在分类器进行机器学习时会“偏袒”多数类样本，使得整体分类准确率大幅度提高，降低了少数类样本的识别率，容易把坏的分成好的。

2.1. 随机欠采样法

随机欠采样法是一种随机抽样的方法，步骤如下：

1) 从不平衡数据集中抽取多数类样本。

2) 然后利用随机数打乱多数类样本。

3) 最后随机抽取一定量的多数类样本与少数类样本进行混合形成新的平衡数据集。一般抽取的数量等于少数类样本的数量[15]。

2.2. Border Line-Smote 过采样法

Border Line-Smote (BLS)是以 Smote 方法为基础发展起来的一种具有人工合成思想的过采样法。在使用 BLS 方法时首先观察少数类样本附近的情况，然后在边界上的两个少数类数据之间随机合成一个新的样本数据，反复如此，不断地增加少数样本的数量，直到形成平衡数据集为止。步骤如下：

1) 对不平衡数据集中的少数类样本重新分类并命名为 X 样本，将 X 样本周围一半以上的多数类样本视为边界上的样本，样本周围均为多数样本视为噪音；

2) 然后利用欧氏距离确定每个样本点与所有训练样本的距离，再根据欧式距离判断其 m 近邻，具体公式如下所示：

$$\rho=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2} \tag{1}$$

3) 根据 m 近邻的样本属性划分样本类，当 m 近邻中有一半以上均为少数类样本划分为安全样本；当 m 近邻中有一半以上均为多数类样本划分为边界样本；当 m 近邻中全为多数类样本时划分为噪音。

4) 计算边界样本与少数类样本之间的 K 近邻，再根据采样倍率 R，选择 P 个 K 近邻与边界样本进行线性插值，采用人工合成的方法产生新的少数类样本，具体公式如下：

$$New\left(Y_i\right)=Y_i+dif_i\times\alpha \tag{2}$$

公式中的 $New\left(Y_i\right)$ 表示新产生的样本，Y_i 表示 X 样本中的多数类，dif_i 表示原样本 Y_i 的 K 近邻，α 表示[0, 1]之间的随机数。

5)将合成的少数类样本与原来的数据集中的训练样本混合形成一个新的训练样本。

2.3. 自适应综合过采样(ADASYN)法

ADASYN 是 He 等人提出的一种过采样方法[16]，ADASYN 最大的优点在于能够自己决定合成多少少数类样本的数量，可以防止过拟合问题的发生[17]。算法步骤如下：

1) 计算不平衡程度

将少数类样本记为 m_s，多数类样本为 m_l，不平衡程度为 $d=m_s/m_l$，其中 $d\in(0,1]$。

2) 计算需要合成的少数类样本的数量

$Y=(m_l-m_s)\times b, b\in[0,1]$，当$b=1$时，$Y$为多数类样本数量和少数类样本数量的差值。

3) 用欧式距离计算每个少数类样本的K个近邻，Ω为K个近邻中属于多数类的样本数目，记比例η为$\eta=\Omega/k, \eta\in[0,1]$，得到每个少数类样本的$\eta_i$。

4) 计算每个少数类样本周围的多数类样本的分布情况，公式如下：

$$\hat{\eta}_i=\eta_i\Big/\sum_{i=1}^{m}\eta_i \tag{3}$$

5) 计算每个少数样本需要合成的数目，计算公式为：$y_i=\hat{\eta}_i*Y$。

6) 在每个待合成的少数类样本周围K近邻中选择一个少数类样本，进行合成样本，直到满足步骤5计算出的样本数目为止。

3. 逻辑回归模型介绍

逻辑回归是Berkson提出的一种常用的分类学习方法[18]，被广泛应用于预测和寻找影响因变量的因素。在二分类中利用逻辑回归将目标值的取值范围规划到0至1之间，并且为了达到损失函数收敛的目的不断用牛顿法或者梯度下降法进行迭代。逻辑回归函数形式：

$$Y_{(v)}=\frac{1}{1+\mathrm{e}^{-v}} \tag{4}$$

$$v=\theta^{\mathrm{T}}Z=\theta_0Z_0+\theta_1Z_1+\theta_2Z_2+\cdots+\theta_nZ_n=\sum_{i=0}^{n}\theta_iZ_i \tag{5}$$

构造预测函数为：

$$h_{\theta(Z)}=g_{(\theta^{\mathrm{T}}Z)}=\frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}Z}} \tag{6}$$

似然函数：

$$L_\theta=\prod_{i=1}^{m}p(y_iz_i;\theta)=\prod_{i=1}^{m}\left[h_{\theta(Z_i)}\right]^{y_i}\left[1-h_{\theta(Z_i)}\right]^{1-y_i} \tag{7}$$

对数似然函数：

$$\begin{aligned}l_{(\theta)}&=\log L(\theta)=\sum_{i=1}^{n}\left[y_i\log h_{\theta(Z_i)}+(1-y_i)\log\left(1-h_{(Z_i)}\right)\right]\\&=\sum_{i=1}^{n}\left[y_i(\theta*z_i)-\log\left(1+\exp(\theta*z_i)\right)\right]\end{aligned} \tag{8}$$

逻辑回归模型：

$$P(Y=1|Z)=\frac{\mathrm{e}^{\bar{\theta}Z}}{1+\mathrm{e}^{\bar{\theta}Z}};P(Y=0|Z)=\frac{\mathrm{e}^{\bar{\theta}Z}}{1+\mathrm{e}^{\bar{\theta}Z}} \tag{9}$$

公式中的θ为权重向量，$\theta*Z$是θ和Z的内积，对$l_{(\theta)}$求最大值，得到θ的估计值。

4. 数据处理及实验结果分析

4.1. 数据处理

本实验数据来自于和鲸社区一个已经脱敏的银行客户信用卡数据集，其中有1个目标值，10个特征值，共150,000条数据，经过对缺失值处理之后剩下11万余条数据，数据集的特征如表1所示。

Table 1. Data set feature table
表 1. 数据集特征表

目标和特征值名称	含义	均值	标准偏差	变量赋值说明
Serious Dlqin2yrs	是否逾期	无	无	0：未违约 1：违约
Revolving Utilization of Unsecured Lines	信用卡和个人信贷额度的总余额	6.05	260.67	实际值
age	年龄	50.42	13.49	实际值
Number of Time 30~59 Days Past Due Not Worse	过去两年借款人逾期30~59天次数	0.26	0.72	实际值
Debt Ratio	负债率	26.41	386.111	实际值
Monthly Income	月收入	6711.53	11,434.77	实际值
Number of Open Credit Lines and Loans	未偿还贷款量和信贷额度	8.80	5.17	实际值
Number of Times 90 Days Late	借款人逾期90天及以上次数	0.09	0.43	实际值
Number Real Estate Loans or Lines	抵押贷款和房地产贷款数量	1.07	1.15	实际值
Number of Time 60~89 Days Past Due Not Worse	过去两年借款人逾期60~89天次数	0.07	0.33	实际值
Number of Dependents	家属人数	0.87	1.15	实际值

由表 1 可知信用卡和个人信贷额度的总余额、年龄、负债率、月收入的标准偏差过大，说明它们的指标数值相差过大需要进行标准化，我们直接通过 SPSS 软件中相关的功能进行标准化，标准化之后的均值为 0，标准偏差为 1。

目标值表现的是客户是否违约，其中有 8136 名客户违约、108,774 名客户未违约，未违约客户和违约客户的比例如图 1 所示，数据集存在着严重的不平衡性，所以需要对数据集中的数据进行平衡处理。

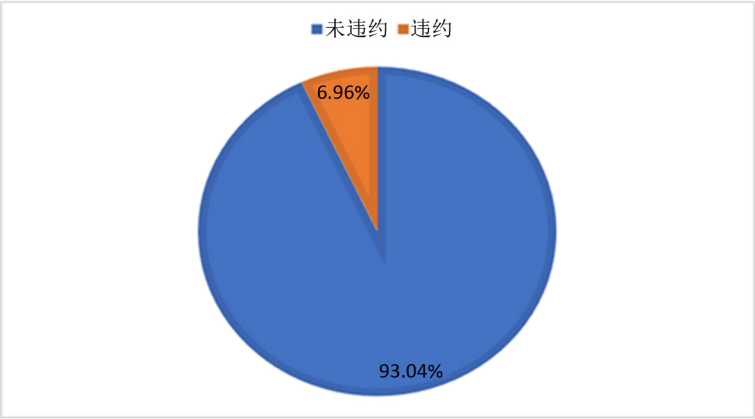

Figure 1. Customer type pie chart
图 1. 客户类型扇形图

4.2. 实验结果及分析

首先通过 Python 分别经过随机欠采样法、Border Line-Smote 过采样法、自适应综合过采样法(ADASYN)对数据集进行平衡化处理，然后对平衡后的数据集分别用逻辑回归模型进行分类，70%的数据作为训练集，30%的数据作为预测集，分类之后得出的指标如表 2 所示。因为是比较三种方法的平衡数据的能力所以不需要设置样本权重参数；类型权重参数选择 balanced 让类库自行计算类型权重，计算方法如下：

$$w = n_s \Big/ \left(n_c * n_p \right) \tag{10}$$

其中 n_s 代表样本数， n_c 代表类别数量， n_p 代表输出每个类的样本数。

Table 2. Three sampling methods are used in the logistic regression model for the results of the indicators
表 2. 三种采样方法在逻辑回归模型结果的指标

平衡方法	accuracy	precision	recall	G-mean	f-score
随机欠采样	0.7351	0.6855	0.8687	0.7229	0.7663
BLS	0.8089	0.7709	0.8790	0.8058	0.8214
ADASYN	0.6988	0.6647	0.8032	0.6908	0.7274

accuracy 代表分类器对数据分类的准确率：

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

precision 代表着分类器对多数类样本正确分类的能：

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

recall 表示分类器对少数类样本正确分类的能力：

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

G-mean 代表着分类器对多数类和少数类分类的整体能力：

$$\text{G-mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} * \frac{\text{TN}}{\text{TN} + \text{FP}}} \tag{14}$$

f-score 表示了分类器的稳定性：

$$\text{f-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{15}$$

以上公式中 TP 是真正例，表示客户为守信，分类器判断为守信；TN 是真反例，表示客户为违约，分类器判断为违约；FP 是假正例，表示客户为守信，分类器判断为违约；FN 是假反例，表示客户为违约，分类器判断为守信。

accuracy、precision、recall、G-mean、f-score 的值越高就代表了分类器对数据处理的性能越强，预测能力越准。明显利用 BLS 方法平衡过的数据经过逻辑回归模型处理后产生的五项数据均比其它两种方法优异，五项数据中的 accuracy、recall、G-mean、f-score 均超过了 0.8，其中 recall 更是达到了 0.879；accuracy 比随机欠采样方法得到的 accuracy 高出 0.0638，更比 ADASYN 方法得到的 accuracy 高出 0.1；precision 比其它两种方法得到的 Precision 至少高出近 0.09；ADASYN 方法平衡过的数据经过逻辑回归模型处理之后各项数据最低；随机欠采样方法得出的结果介于 BLS 和 ADASYN 之间。

图 2~4 分别为随机欠采样法、BLS 方法、ADASYN 方法得出的 ROC 曲线图，ROC 曲线反映了敏感性和特异性连续变量的综合指标，用构图法揭示了敏感性和特异性的相互关系，曲线下面积越大诊断准确性越高，即 ROC 曲线越靠向左上角说明分类器的分类结果越可靠。我们可以观察到用逻辑回归模型处理 BLS 方法平衡之后的数据集得到的 ROC 曲线最靠近左上角，处理随机欠采样方法平衡之后的数据集得到的 ROC 曲线次之，处理 ADASYN 方法平衡之后的数据集得到的 ROC 曲线相比较而言最靠近右边。

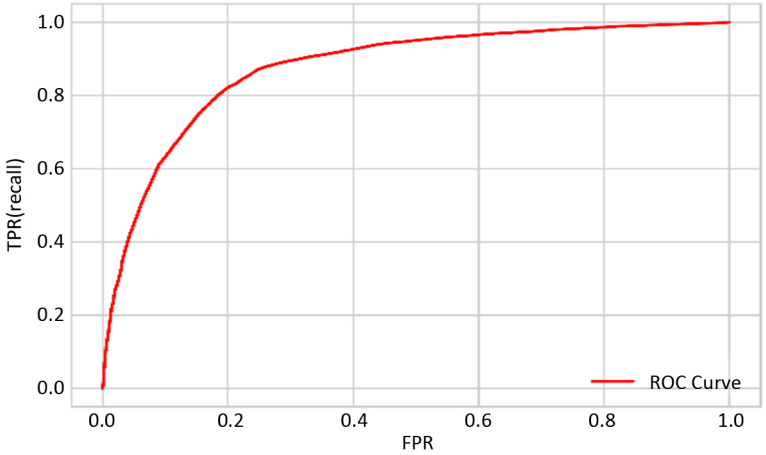

Figure 2. ROC curve for the random undersampling method
图 2. 随机欠采样法 ROC 曲线图

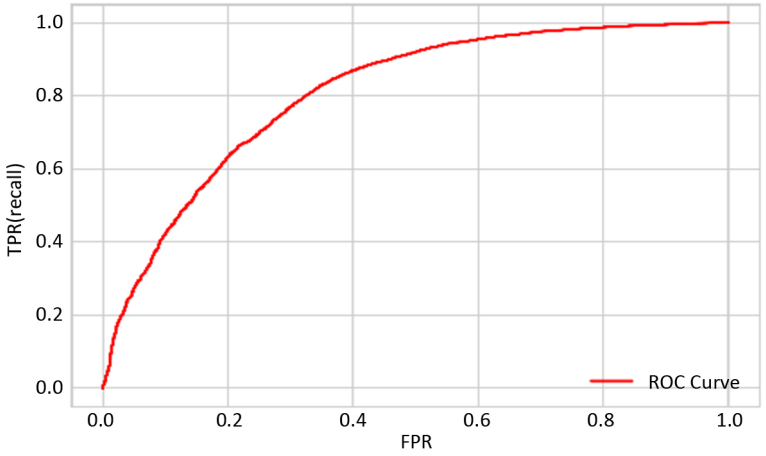

Figure 3. ROC curve of BLS method
图 3. BLS 法 ROC 曲线图

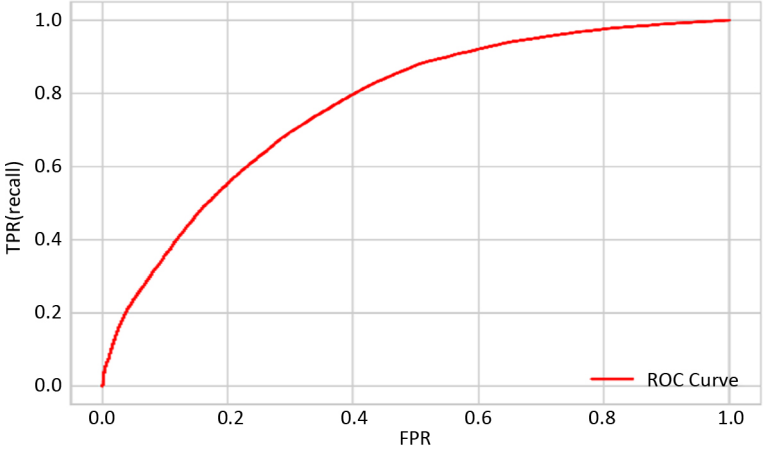

Figure 4. ROC curve of ADASYN method
图 4. ADASYN 法 ROC 曲线图

综上所述利用 BSL 方法处理不平衡数据集更适合逻辑回归模型。

5. 总结与展望

通过第四章实验的结果及分析可知，经过 BLS 方法平衡数据集之后再利用逻辑回归模型进行分类得出的各项数据均比通过随机过采样法和 ADASYN 方法平衡数据集之后再利用逻辑回归模型进行分类的各项数据更加优异，与其它两种方法相比 BLS 方法更适合逻辑回归模型下的不平衡数据集处理。同时我们可以得知用同一种分类器、不同的数据平衡方法得出的各项分类数据是不同的。

金融机构对客户进行准确的信用评级可以提高对风险的把控能力，然而信用数据往往是不平衡的，所以在信用评级之前需要平衡数据集。本文只是研究随机欠采样法、BLS 过采样法、ADASYN 过采样法等三种方法中何种方法更适合逻辑回归模型下的不平衡数据集处理，最终得出 BLS 方法更适合逻辑回归模型下的不平衡数据集处理，在以后我们是否可以深度研究不同的分类器适合何种不平衡数据集处理方法，以此使分类器的准确度最大化，这一问题还有待探究。

基金项目

江苏省社会科学基金项目，信用挖掘为中心人工智能普惠金融服务技术研究，批准号为 18GLA004。

参考文献

[1] 徐丽丽, 闫德勤, 高晴. 基于聚类欠采样的极端学习机[J]. 微型机与应用, 2015(17): 81-84.

[2] Paolo, S. (2010) A Multi-Objective Optimization Approach for Class Imbalance Learning. *Computers in Biology and Medicine*, **40**, 509-518. https://doi.org/10.1016/j.compbiomed.2010.03.005

[3] 王和勇, 樊泓坤, 姚正安, 李成安. 不平衡数据集的分类方法研究[J]. 计算机应用研究, 2008(5): 1301-1303+1308.

[4] 顾东晓, 李培培, 杨雪洁. 网络在线预约挂号系统用户的爽约行为研究[J]. 情报科学, 2017, 35(5): 99-106.

[5] Han, H., *et al*. (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Application Research of Computers*, **56**, 66-68.

[6] Vasu, M. and Ravi, V. (2011) A Hybrid Under-Sampling Approach for Mining Unbalanced Datasets: Applications to Banking and Insurance. *International Journal of Data Mining, Modelling and Management*, **3**, 75-105. https://doi.org/10.1504/IJDMMM.2011.038812

[7] Li, H. and Sun, J. (2012) Forecasting Business Failure: The Use of Nearest-Neighbour Support Vectors and Correcting Imbalanced Samples Evidence from the Chinese Hotel Industry. *Tourism Management*, **33**, 622-634. https://doi.org/10.1016/j.tourman.2011.07.004

[8] Sundarkumar, G.G. and Ravi, V. (2015) A Novel Hybrid Undersampling Method for Mining Unbalanced Datasets in Banking and Insurance. *Engineering Applications of Artificial Intelligence*, **37**, 368-377. https://doi.org/10.1016/j.engappai.2014.09.019

[9] Błaszczyński, J. and Stefanowski, J. (2015) Neighbourhood Sampling in Bagging for Imbalanced Data. *Neurocomputing*, **150**, 529-542. https://doi.org/10.1016/j.neucom.2014.07.064

[10] Bi, J.J. and Zhang, C.S. (2018) An Empirical Comparison on State-of-the-Art Multi-Class Imbalance Learning Algorithms and a New Diversified Ensemble Learning Scheme. *Knowledge-Based Systems*, **158**, 81-93.

[11] Namvar, A., Siami, M., Rabhi, F. and Naderpour, M. (2018) Credit Risk Prediction in an Imbalanced Social Lending Environment. *International Journal of Computational Intelligence Systems*, **11**, 925-935. https://doi.org/10.2991/ijcis.11.1.70

[12] 高阳, 刘其成, 牟春晓. 基于蚁群聚类的不平衡数据过采样方法[J/OL]. 烟台大学学报(自然科学与工程版), 1-8 [2020-11-19].

[13] 蒋华, 江日辰, 王鑫, 王慧娇. ADASYN 和 SMOTE 相结合的不平衡数据分类算法[J]. 计算机仿真, 2020, 37(3): 254-258+420.

[14] Guo, H.X., Li, Y.J., Shang, J., Gu, M.Y., Huang, Y.Y. and Gong, B. (2016) Learning from Class-Imbalanced Data: Review of Methods and Applications. *Expert Systems with Applications*, **73**, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

[15] 宋捷. 不平衡数据处理方法综述[J]. 统计与决策, 2014(3): 100-102.

[16] He, H. and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge & Data Engineering*, **21**, 63-84. https://doi.org/10.1109/TKDE.2008.239

[17] 刘金平，周嘉铭，贺俊宾，唐朝晖，徐鹏飞，张国勇．面向不均衡数据的融合谱聚类的自适应过采样法[J/OL]. 智能系统学报, 1-8. http://kns.cnki.net/kcms/detail/23.1538.TP.20200827.1317.008.html, 2020-10-30.

[18] Berkson, J. (2012) Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, **39**, 357-365. https://doi.org/10.2307/2280041