ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 6,Nov. 2023,pp. 1482－1494
DOI:10. 16337/j. 1004-9037. 2023. 06. 021

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于NaN-Bicluster SMOTE的非均衡信贷数据分类研究及应用

何　亮，徐海燕，陈　璐

（南京航空航天大学经济与管理学院，南京 211106）

摘　要：为了有效评估非均衡信贷数据中的借款人信用风险，基于合成少数过采样技术（Synthetic minority oversampling technique，SMOTE）、自然近邻（Natural neighbor，NaN）和双聚类（Bicluster）构建了NaN-Bicluster SMOTE方法以改进SMOTE。首先使用无参数的自然近邻设定采样样本选取的逻辑规则，规避了r近邻划分样本时产生的不稳定性；其次基于自然近邻稳定结构规定安全范围设定的逻辑规则，避免合成样本成为噪声样本；然后使用双聚类挖掘局部规则，以合成样本继承局部规则的方式改进SMOTE合成公式；最后，在Prosper小额贷款平台的非均衡信贷数据集上将NaN-Bicluster SMOTE与若干采样方法和机器学习模型进行对比分析，并进一步使用统计检验方法验证其性能的优越性。

关键词：小额贷款；信用风险；合成少数过采样技术；自然近邻；双聚类

中图分类号：TP181　　**文献标志码：**A

Research and Application of Imbalanced Credit Data Classification Based on NaN-Bicluster SMOTE

HE Liang, XU Haiyan, CHEN Lu

(College of Economics and Management, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China)

Abstract: To assess borrower's credit risk using imbalanced data, we propose an improved SMOTE, called NaN-Bicluster SMOTE, which is based on synthetic minority oversampling technique (SMOTE), natural neighbor (NaN) and bicluster. Firstly, we use parameterless NaN to set logical rules for sampling sample selection, avoiding the instability caused by r nearest neighbor partitioning of samples. Secondly, based on the neighbor relationship of stable structure, we set logical rules that specify security range to avoid samples becoming noise samples. Then, we use bicluster to mine local rules, synthetic samples inherit local rules, and synthetic formula is improved. Finally, we apply several sampling methods and machine learning models, carry out various experiments of NaN-Bicluster SMOTE and comparative models on Prosper's credit data, and further use statistical testing methods to verify the performance of NaN-Bicluster SMOTE.

Key words: microloans; credit risk; synthetic minority oversampling technique (SMOTE); natural neighbor (NaN); bicluster

基金项目：国家自然科学基金面上项目(71971115)；国家自然科学基金青年项目(72201126)；智能决策与数字化运营工业和信息化部重点实验室项目(NJ2023027)。
收稿日期：2022-07-13；**修订日期：**2022-10-05

引　言

小额贷款作为小额度的持续性信贷形式[1],在缓解个人及中小企业的资金约束、推动经济持续增长的过程中,发挥着不可替代的作用[2],已经成为一种主要的信贷模式。然而,越来越多的小额贷款平台因为借款人的信用风险而蒙受巨大损失,甚至濒临破产和倒闭[3-4],因此,有效甄别信用优质的借款人对信贷行业的健康发展有着深远影响。

机器学习模型可以有效甄别借款人的信用风险,借助于信贷数据对借款人的信用进行“违约”与“非违约”风险的分类评估。决策树(Decision tree,DT)[5]、逻辑回归(Logistic regression,LR)[6]、K近邻(K nearest neighbor,KNN)[7]、支持向量机(Support vector machine,SVM)[8-9]和神经网络(Neural network,NN)[10]等机器学习模型已被广泛用于评估信用风险,并被证实具有较好的性能。而在现实信贷数据中,“违约”借款人占比极少,与“非违约”借款人的数量差异极大,这种差异使得信贷数据具有非均衡特征。机器学习模型在处理非均衡数据时,往往过多地学习“非违约”的多数类样本,对“违约”的少数类样本识别欠缺,致使机器学习模型对“违约”的少数类样本识别能力近乎为零,无法分类出关键的“违约”借款人也就不具备实际应用价值。如何妥善处理信贷数据的非均衡特征,提升机器学习模型的分类能力,成为信用风险评估中亟待解决的问题。

现有研究中,非均衡数据的处理方式可分为成本敏感方法[11]、算法级方法[12]和数据级方法[13]。数据级方法使用最为广泛,包含欠采样方法和过采样方法。两种采样方式虽然同样易受噪声和边界样本的影响,但欠采样方法丢失过多数据,数据原有特性被破坏,而过采样方法的对象是决策者关注的“违约”少数类样本,应用场景较欠采样方式更为广泛。合成少数过采样技术(Synthetic minority oversampling technique,SMOTE)[14]是一种著名的过采样方法,通过在邻近的少数类样本之间进行随机线性插值合成新少数类样本,从而达到数据均衡,被运用于信用风险评估[15]、虚假评论识别[16]及工业故障检测[17]等诸多领域。然而,诸多研究表明SMOTE的合成过程易受噪声和边界样本影响,均衡化后的数据往往不符合预期[18-19],因此,改进合成过程以此提升SMOTE的均衡能力至关重要。

采样方式和合成公式是SMOTE改进的两个主要方向。采样方式的改变是指在指定区域内合成少数类样本。Han等[20]认为处于正负类之间的边界样本被误分的概率较大,于是基于在边界区域合成样本的思想构建了Borderline SMOTE,但其忽视了噪声的影响;Bunkhumpornpat等[21]通过计算安全水平范围,提出了Safe-Level SMOTE,将合成的少数类样本置于安全区域,但合成的少数类样本过于靠近多数类样本,从而干扰了分类过程;赵冬雪等[22]融合采样方式和粗糙集属性约简技术,构建动态集成的分类模型;邢延等[23]使用类别混叠度作为采样的依据,指导非均衡数据的分类;吴志峰等[24]综合使用逐级优化递减欠(Optimization of decreasing reduction,ODR)采样算法和数据清洗方法,构建了融合自适应变核参数模型(ODR-BSMOTE-TOMEK adaptive support vector machine,OBT-Adaptive-SVM),以提升分类性能。改变合成公式旨在弥补线性随机插值带来的噪声样本过度泛化且随机性较强的不足。Li等[25]通过计算少数类样本与其自然近邻(Natural neighbor,NaN)样本的属性差值,替代传统SMOTE的属性差值计算,使得合成样本与少数类样本具有更多的相似性,并消减噪声样本的影响;Soltanzadeh等[26]构建了边界约束的合成少数过采样技术(Range-controlled synthetic minority oversampling technique,RCSMOTE),通过计算属性的范围,融合差分计算并设计了一套全新的合成公式,但对高维属性的数据运算较为缓慢。

现有研究对传统SMOTE进行了多方改进,但采样样本选取、合成公式制定仍存在部分欠缺之处,具体如下:(1)现有研究通过r近邻(r-nearest neighbor,NNr)划分安全样本、边界样本和噪声样本,以此选定边界样本进行采样。然而,该种划分方式严重依赖于近邻参数r的设定,整个合成过程对参数取值

极为敏感,存在较大的不稳定性。(2)在边界样本的少数类样本和其r近邻少数类样本合成新样本时,无法保证少数类样本与r近邻少数类样本同时处于安全范围,存在合成样本置于多数类范围内的现象,成为噪声样本并混淆了分类边界。(3)随机线性差值的合成公式忽视了合成样本与少数类样本之间的关联规则,合成样本应继承少数类样本的规则,以体现两者的相关性。

针对现有研究中存在的问题,本文在SMOTE基础上,借助于无参数设定并提供稳定结构的自然近邻、用以挖掘数据隐藏的局部信息及表达同类数据关联规则的双聚类(Bicluster)[27-28],构建了NaN-Bicluster SMOTE,以此从采样样本选取、安全范围设定和合成公式制定多个角度改进SMOTE。首先,使用无参数的自然近邻设定采样样本选取的逻辑规则,找出可疑样本并用离群度对可疑样本进行边界样本和噪声样本的甄别,取代依赖于参数设定的r近邻样本划分方式,更为客观地选取参与采样的边界样本;其次,对于边界样本中的少数类样本,借助自然近邻的稳定结构规定安全范围设定的逻辑规则,规避合成样本置于多数类范围的情形;然后,在安全范围内,通过双聚类提取局部规则,使得合成样本继承少数类样本的局部规则,从而实现合成公式的改进;最后,结合Prosper小额贷款平台的信贷数据,与已有的多种采样方法和机器学习模型进行大量的对比分析,并使用统计检验方法验证本文所提的NaN-Bicluster SMOTE方法在信贷数据上的均衡性能。

1 理论分析

1.1 SMOTE

(1)基础理论

SMOTE通过少数类样本$x_i^{\min}$、$x_j^{\min}$的随机线性插值合成新少数类样本x_{new},实现数据的均衡化[20],合成公式可表示为

$$x_{\text{new}}=x_i^{\min}+\text{gap}\times(x_i^{\min}-x_j^{\min}) \tag{1}$$

式中gap表示0~1之间的随机数。对于任意一个少数类样本$x_i^{\min}$,挑选其r近邻集合中的少数类样本$x_j^{\min}$,通过式(1)合成新的少数类样本x_{new}。

(2)理论不足

现有研究中SMOTE的采样样本选取、合成公式制定仍有待完善,具体问题可归纳为以下3点:

第一,采样样本的选取过程严重依赖于参数设定。现有研究通过r近邻集合中少数类样本的个数r',实现安全样本$S(r/2<r'\leqslant r)$、边界样本$B(0<r'\leqslant r/2)$和噪声样本$N(r'=0)$的划分,并对边界样本进行过采样,如图1(a)所示。但该过程依赖近邻参数r的设定,不同r得到的边界样本集合是不同的,使SMOTE的性能存在较大不稳定性。

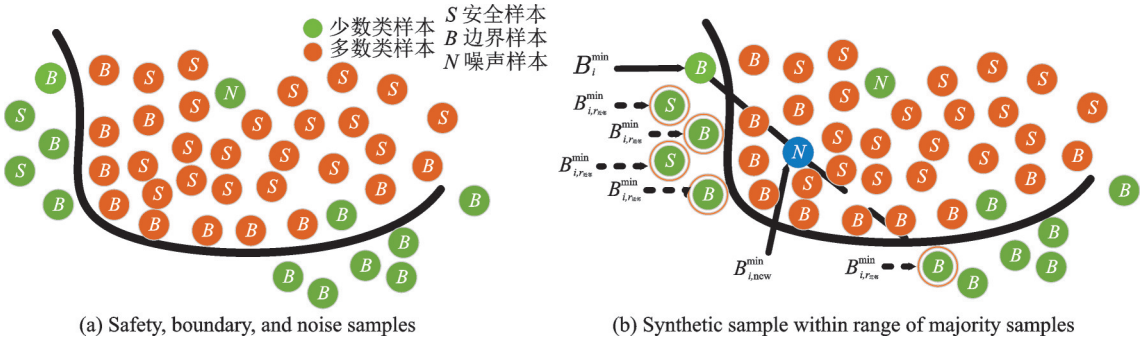

图1 SMOTE采样样本选取及安全范围示意图

Fig.1 Schematic diagram of sample selection and safe range setting for SMOTE sampling

第二,无法保证合成过程处于安全范围。在图1(b)中,$B_i^{\min}$为边界样本中的少数类样本,$B_{i,r_{近邻}}^{\min}$为$B_i^{\min}$的r近邻集合中的少数类样本,在$B_i^{\min}$和$B_{i,r_{近邻}}^{\min}$之间合成新少数类样本$B_{i,\text{new}}^{\min}$,但参与采样的样本$B_i^{\min}$与$B_{i,r_{近邻}}^{\min}$未处于安全合理的范围,使得合成样本$B_{i,\text{new}}^{\min}$处于多数类样本的范围之内,成为不符合预期的噪声样本。

第三,合成公式的制定有较大提升空间。式(1)中的$x_i^{\min}$体现合成样本与原样本的相似性,gap$\times(x_i^{\min}-x_j^{\min})$则体现了二者的差异性,在一定程度上权衡了相似性和差异性,但尚未充分体现$x_i^{\min}$、$x_j^{\min}$和x_{new}之间的内在关联规则。

1.2　自然近邻

自然近邻相较于r近邻,是一种无参数设定的检测技术[29],被广泛运用于异常值检测[30]和分层聚类[31]等领域;其原理是不断搜索每个样本的r近邻,直到每一个样本都有一个相互的r近邻,形成稳定结构,此时自然近邻的特征值λ等于r[32],且每个样本的自然近邻集合的元素个数均不同,可以表述为

$$(\forall x_i)(\exists x_j)(r\in n)\wedge(x_i\neq x_j)\rightarrow(x_i\in NN_r(x_j))\wedge(x_j\in NN_r(x_i)) \tag{2}$$

式中:x_i和x_j分别表示两个不同的样本;$NN_r(x_i)$表示x_i的r近邻集合;$NN_r(x_j)$表示x_j的r近邻集合。在一个稳定结构中,对于任意一个x_i,都存在与之不同的x_j,使得x_i与x_j互为r近邻。

自然近邻的定义如式(3)所示,被看作为稳定结构中构成的友邻关系

$$x_i\in \text{NaN}(x_j)\Leftrightarrow x_j\in NN_\lambda(x_i)\wedge x_i\in NN_\lambda(x_j) \tag{3}$$

式中$\text{NaN}(x_j)$表示x_j的自然近邻集合。x_i属于x_j的自然近邻集合的充要条件是,在稳定结构中,x_j是x_i的λ近邻,且x_i是x_j的λ近邻。

1.3　双聚类

相较于传统聚类搜寻的全局信息,双聚类挖掘了数据隐藏的局部信息,用以表达同一类数据的关联规则[27]。其原理是同时对矩阵的行和列进行聚类,寻找具有紧密联系的子矩阵,即为双聚类结果,如图2所示。

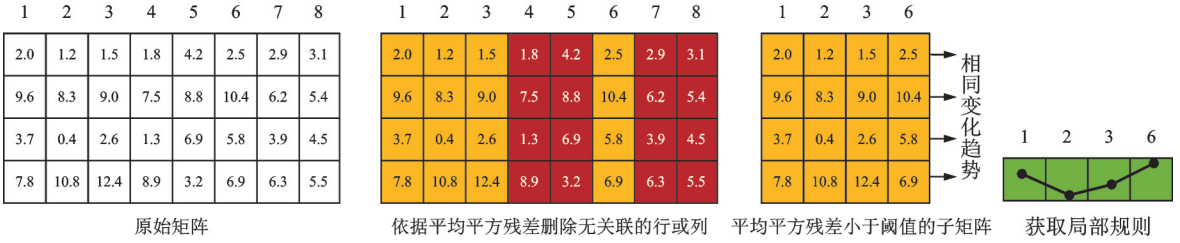

图2　双聚类示意图

Fig.2　Schematic diagram of bicluster

在图2中,依据平均平方残差删除无关联的行或者列,得到平均平方残差在阈值范围内的子矩阵,进而归纳出局部规则,具体定义如下[33]。$m\times n$的数据集构成原始矩阵$\boldsymbol{A}$,G为m个样本的集合,C为n个属性的集合,x_{ij}为第i个样本的第j个属性值。双聚类结果为Bicluster$=(I,J)$,其中I为G的子集,J为C的子集,子矩阵Bicluster的平均平方残差$H(I,J)$的计算公式为

$$H(I,J)=\frac{1}{|I||J|}\sum_{i\in I,j\in J}\left(x_{ij}-x_{Ij}-x_{iJ}+x_{IJ}\right)^2 \tag{4}$$

式中：x_{iJ}表示Bicluster第i行的平均值，$x_{iJ}=\frac{1}{|J|}\sum_{j\in J}x_{ij}$；$x_{Ij}$表示Bicluster第$j$列的平均值，$x_{Ij}=\frac{1}{|I|}\sum_{i\in I}x_{ij}$；$x_{IJ}$表示Bicluster的平均值。通过设定子矩阵Bicluster的最大平均平方残差δ，决定Bicluster中行列的添加与删除，$\delta\in(0,1)$。δ越小，允许的最大平均平方残差越小，表明子矩阵的内在联系越紧密，则Bicluster和局部规则的规模越小；反之，Bicluster和局部规则的规模越大[34-35]。

2 NaN-Bicluster SMOTE方法构建

为完善SMOTE的理论不足，构建了NaN-Bicluster SMOTE方法的框架结构，依据该框架结构依次介绍本文工作。

2.1 NaN-Bicluster SMOTE框架构建

基于自然近邻、双聚类和SMOTE，构建了NaN-Bicluster SMOTE方法，分为采样样本选取、安全范围设定和局部规则提取与合成公式改进3个阶段，具体如图3所示。

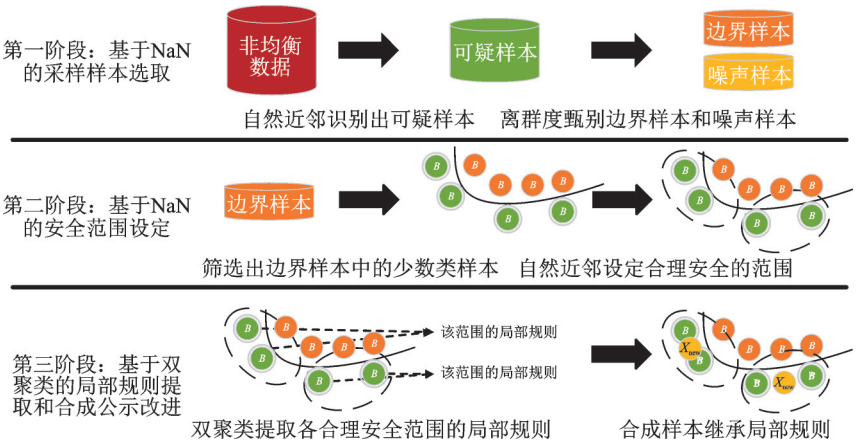

图3 NaN-Bicluster SMOTE方法框架图

Fig.3 Framework of NaN-Bicluster SMOTE

第一阶段，采样样本选取。使用无参数的自然近邻设定逻辑规则，得出非均衡数据中的可疑样本。设定离群度阈值，甄别出可疑样本中的边界样本和噪声样本，删除噪声样本减少其造成的不利影响，选取边界样本参与采样。替代了r近邻的样本划分方式，增强了模型输出结果的稳定性。

第二阶段，安全范围设定。使用自然近邻的稳定结构设定逻辑规则，为边界样本中的少数类样本设定合理安全的范围，避免合成样本处在多数类样本范围内、成为扰动SMOTE均衡性能的噪声样本。

第三阶段，局部规则提取与合成公式改进。在安全范围内，使用双聚类提取各个范围内少数类样本的局部规则，充分体现安全范围内少数类样本的共有特性。通过合成样本继承局部规则的方式，实现SMOTE合成公式的改进。

2.2 基于自然近邻的采样样本识别

使用自然近邻设定可疑样本识别的逻辑规则。若样本自然近邻中存在与其分类标签不一致的样本，则该样本称为可疑样本(Suspicious sample，SE)；若样本的标签与其所有自然近邻标签完全一致，则该样本称为正常样本(Normal sample，NE)，其定义分别如式(5)和式(6)所示。

$$x_i\in \mathrm{SE}\Leftrightarrow(\exists x_j)(x_j\in \mathrm{NaN}(x_i))\wedge(l(x_i)\neq l(x_j)) \tag{5}$$

$$x_i \in \mathrm{NE} \Leftrightarrow (\forall x_j)(x_j \in \mathrm{NaN}(x_i)) \wedge (l(x_i) = l(x_j)) \tag{6}$$

在式(5)中:对于x_i而言,自然近邻中存在x_j的标签$l(x_j)$与自身标签$l(x_i)$不一致,则x_i属于可疑样本SE;式(6)中:对于x_i而言,自然近邻中不存在任何一个x_j的标签$l(x_j)$与自身标签$l(x_i)$不一致,则x_i属于正常样本NE。

使用自然近邻和离群值设定边界样本选取的逻辑规则。通过设定离群度甄别得出边界样本和噪声样本,可分别表示为

$$x_i \in N \Leftrightarrow (\forall x_j)(x_i \in \mathrm{SE}) \wedge (x_j \in \mathrm{NaN}(x_i)) \wedge (l(x_i) \neq l(x_j)) \tag{7}$$

$$x_i \in B \Leftrightarrow (\exists x_j)(x_i \in \mathrm{SE}) \wedge (x_j \in \mathrm{NaN}(x_i)) \wedge (l(x_i) = l(x_j)) \tag{8}$$

对于可疑样本SE中的样本x_i而言,如果x_i的自然近邻集合$\mathrm{NaN}(x_i)$中任意一个样本x_j的标签$l(x_j)$与其标签$l(x_i)$不一致时,即x_i的友邻关系全为异类样本,记x_i为噪声样本N;如果x_i的自然近邻集合$\mathrm{NaN}(x_i)$中存在一个样本x_j的标签$l(x_j)$与其标签$l(x_i)$一致时,即x_i处于两类样本的边界区域,记x_i为边界样本B。

2.3　基于自然近邻的安全范围设定

现有研究通常选取边界样本中的少数类样本与其r近邻集合中的少数类样本合成样本,但该方式并未考虑到两者是否处于同一安全范围。若两者不处于同一安全范围内,则合成的少数类样本极有可能偏离少数类群体,成为噪声样本。自然近邻在识别可疑样本的同时,提供了一个稳定结构,该稳定结构中的友邻关系可以较好地形成安全范围,规避上述问题。NaN-Bicluster SMOTE安全范围逻辑规则的设定如下所示。

从边界样本B中选出少数类样本$B^{\min}$,可表示为

$$x_i \in B^{\min} \Leftrightarrow (x_i \in B) \wedge (l(x_i) = l^{\min}) \tag{9}$$

式中:对于边界样本B中的样本x_i,若其标签$l(x_i)$与少数类标签$l^{\min}$一致,则x_i属于边界样本中的少数类样本$B^{\min}$。

通过自然近邻稳定结构中的友邻关系,将$B^{\min}$采样过程中的安全范围设定为

$$(\forall x_i)(\forall x_j)(x_i \in B^{\min}) \wedge (x_j \in B^{\min}) \wedge (x_j \in \mathrm{NaN}(x_i)) \tag{10}$$

式中:对于少数类样本$B^{\min}$中的任意一个参与采样的样本x_i,与之配对的样本x_j需同时满足:x_j属于x_i的自然近邻集合$\mathrm{NaN}(x_i)$,且x_j属于$B^{\min}$。将x_i与x_j限定于自然近邻的稳定结构中,从而设定采样的安全范围。

2.4　基于双聚类的局部规则提取和合成公式改进

以双聚类提取出的局部规则为基础,以继承和差异的思想改进合成公式,具体如下所示,设局部规则$R=(R_{\mathrm{attribute}}, R_{\mathrm{value}})$,将$R$映射为属性集$R_{\mathrm{attribute}}$和属性取值集$R_{\mathrm{value}}$,$R_{\mathrm{attribute}}$表示局部规则包含的属性集合,$R_{\mathrm{value}}$表示局部规则包含的属性取值集合。

首先,对于参与采样的边界样本x_i,选取处于安全范围内的样本x_j,计算得出x_i、x_j和自然近邻集合$\mathrm{NaN}(x_i)$的局部规则R,通过式(11)选取差值diff,即

$$\mathrm{diff}^d = \begin{cases} R_{\mathrm{value}} & d \in R_{\mathrm{attribute}} \\ x_i^d - x_j^d & d \notin R_{\mathrm{attribute}} \end{cases} \tag{11}$$

式中:diff^d表示第d个属性的差值,$\mathrm{diff}=\left(\mathrm{diff}^1, \mathrm{diff}^2, \cdots, \mathrm{diff}^d, \cdots\right)$,其含义为,如果样本第$d$个属性属于局部规则属性集$R_{\mathrm{attribute}}$,那么差值$\mathrm{diff}^d$为该属性对应的局部规则属性取值集$R_{\mathrm{value}}$;如果样本第$d$个属性不属于局部规则属性集$R_{\mathrm{attribute}}$,那么差值为传统SMOTE的差值$x_i^d - x_j^d$。

其次,SMOTE合成公式设定为

$$x_{\text{new}} = x_i + \text{gap} \times \text{diff} \quad (12)$$

式中:通过继承x_i、x_j及自然近邻集合的局部规则,使得x_{new}的部分属性取值与x_i、x_j保持一致的趋势,更贴近该安全范围内的样本;通过$x_i^d - x_j^d$随机更改属性取值,使得x_{new}的其余属性取值与x_i、x_j存在一定的差异,增强了多样性,如图4所示。

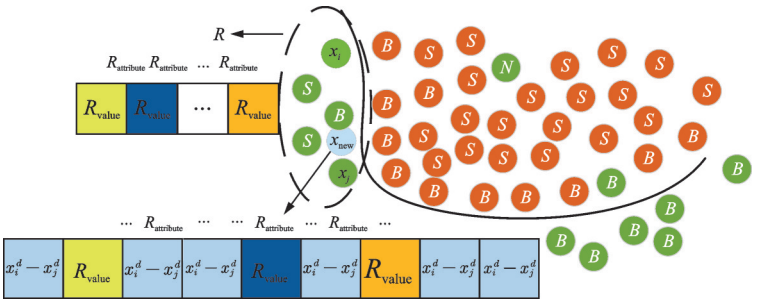

图4 基于双聚类的合成公式改进示意图

Fig.4 Schematic diagram of improved synthetic formula based on bicluster

3 实验和结果分析

为了验证NaN-Bicluster SMOTE方法在非均衡信贷数据上的均衡能力,在Prosper小额贷款平台的真实信贷数据集上,选取SMOTE、Borderline SMOTE、Safe-Level SMOTE和OBT-Adaptive-SVM这4种采样方法,结合DT、LR、KNN、SVM和NN这5种机器学习模型展开对比分析,并采用多种维度的评价准则和统计检验方法评估验证NaN-Bicluster SMOTE方法的性能。

3.1 数据来源

选取全球规模前十的Prosper小额贷款平台的信贷数据[36],该数据包含2013年的2 343条数据与51个属性,基本信息如表1所示。

记Completed状态样本为非违约样本,Charged Off和Defaulted状态样本为违约样本。由于数据若

表1 数据集的基本信息

Table 1 Basic information of data sets

序号	属性	序号	属性	序号	属性
1	贷款原因	18	Prosper评级	35	开立的循环账户数量
2	借款人居住地	19	估计损失	36	债务收入比
3	借款人职业	20	估计有效收益	37	循环账户每月还款数量
4	受雇佣状态	21	预期回报	38	当前拖欠额度
5	受雇时间	22	上市资金的百分比	39	当前拖欠美元数
6	是否为户主	23	收益率	40	循环账户中的美元
7	是否在组群中	24	投资人朋友投资数量	41	使用可用循环信用的百分比
8	年收入范围	25	投资人朋友投资数	42	通过银行卡的可用信用总额
9	支持其收入必要文件	26	提供资金投资者数量	43	开设的交易行数
10	贷款期限	27	信用评分	44	从未拖欠的交易数量
11	借款人年利率	28	信贷资料被扣除的日期	45	最近6个月内开设交易数量
12	贷款标利率	29	评分范围上限	46	已经向Prosper借入的资金
13	客户月收入	30	评分范围下限	47	最近6个月查询征信次数
14	日期	31	第一条信用额度建立日期	48	征信记录查询的总次数
15	贷款起源的季度	32	当前信贷额度	49	过去7年违约次数
16	预定的每月贷款支付	33	未清信用额度	50	过去10年的公共记录数量
17	投资人的建议数量	34	过去七年中的信用额度	51	过去12个月的公共记录数量

干属性取值为文本型分类数据,机器学习模型难以学习该类数据,本文以有序排列的1、2、3等数值型数字代替文本型分类数据,将文本型分类数据转化为数值型分类数据。例如,“借款人居住地”包含48种文本型取值,分别使用1、2、…、48的数值进行对应,对应顺序不影响后期分类结果。对预处理后的数据进行标准化处理。数据处理后,违约样本数为274,非违约样本数为2 069,非均衡比约为1∶7。

3.2　模型评价准则的选取

为了公平比较NaN-Bicluster SMOTE与其他采样方法在非均衡数据上的性能差异,选取多种维度的评价准则进行性能评估。首先,选取G-mean这一非均衡数据的评价准则,计算公式为[37-38]

$$\text{G-mean}=\sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}}\times\frac{\text{TN}}{\text{TN}+\text{FP}}} \tag{13}$$

式中:被正确分类的正类样本(True positive)记为TP,被正确分类的负类样本(True negative)记为TN,将负类样本错误分类为正类样本(False positive)记为FP,将正类样本错误分类成负类样本(False negative)记为FN。TP/(TP+FN)表示实际为正类的样本被正确分类的比例,TN/(TN+FP)表示实际为负类的样本被正确分类的比例,G-mean为两种比例的几何平均数,同时考虑了正负类样本的比例,适用于非均衡数据集性能的评估。其次,受试者工作特征曲线下与坐标轴围成的面积(Area under curve,AUC)[39]被广泛用以评价模型的分类性能,因此选取其为非均衡数据的评价准则。

3.3　实验结果分析

结合机器学习模型中的DT、LR、KNN、SVM和NN这5种分类器,将构建的NaN-Bicluster SMOTE与现有研究中的SMOTE、Borderline SMOTE和Safe-Level SMOTE、OBT-Adaptive-SVM进行对比分析。采用交叉验证得到NaN-Bicluster SMOTE与对比模型在G-mean和AUC上的均衡能力。鉴于试错法得到5折交叉验证和10折交叉验证差异不显著,遂选取效率更高、运算更快的5折交叉验证用以实验;根据信贷数据1∶7的非均衡比例和现有研究的取值[25-26],均衡比例k分别取值为3、5和7;依据平均平方残差的定义和现有研究的取值[27-28]分别设δ为0.01、0.10和0.30。实验结果如表2~4所示。

从表2中可以得出,当非均衡比例$k=3$时,除了Safe-Level SMOTE+NN在G-mean上表现较优,Safe-Level SMOTE+KNN、Borderline SMOTE+NN、Safe-Level SMOTE+NN在AUC上表现较优之外,NaN-Bicluster SMOTE的均衡能力均为最优。这在一定程度上说明了考虑合成范围设定、以局

表2　NaN-Bicluster SMOTE与对比模型的实验结果(k=3)

Table 2　Experimental results of NaN-Bicluster SMOTE and contrast models (k=3)

模型	评价准则	无采样模型参与	SMOTE	Borderline SMOTE	Safe-Level SMOTE	NaN-Bicluster SMOTE		
						$\delta=0.01$	$\delta=0.10$	$\delta=0.30$
DT	G-mean	0.000 0	0.830 1	0.856 0	0.852 4	0.902 4	0.902 5	0.878 2
	AUC	0.500 0	0.880 0	0.880 0	0.890 0	0.920 0	0.920 0	0.910 0
LR	G-mean	0.501 0	0.567 2	0.749 0	0.802 3	0.911 4	0.912 4	0.914 6
	AUC	0.760 0	0.680 0	0.810 0	0.870 0	0.930 0	0.930 0	0.930 0
KNN	G-mean	0.467 6	0.808 3	0.773 9	0.789 5	0.820 4	0.825 8	0.822 3
	AUC	0.570 0	0.810 0	0.820 0	0.890 0	0.820 0	0.830 0	0.830 0
SVM	G-mean	0.000 0	0.847 6	0.754 0	0.808 4	0.939 1	0.939 1	0.939 1
	AUC	0.580 0	0.680 0	0.810 0	0.860 0	0.870 0	0.870 0	0.890 0
NN	G-mean	0.486 0	0.811 4	0.890 8	0.905 9	0.824 4	0.822 1	0.825 7
	AUC	0.640 0	0.870 0	0.910 0	0.910 0	0.870 0	0.870 0	0.870 0

表3 NaN-Bicluster SMOTE与对比模型的实验结果(k=5)

Table 3 Experimental results of NaN-Bicluster SMOTE and contrast models (k=5)

模型	评价准则	无采样模型参与	SMOTE	Borderline SMOTE	Safe-Level SMOTE	NaN-Bicluster SMOTE		
						$\delta=0.01$	$\delta=0.10$	$\delta=0.30$
DT	G-mean	0.000 0	0.881 2	0.848 8	0.881 1	0.916 9	0.908 6	0.921 0
	AUC	0.500 0	0.910 0	0.910 0	0.890 0	0.940 0	0.940 0	0.940 0
LR	G-mean	0.501 0	0.551 3	0.756 6	0.780 8	0.922 7	0.923 6	0.921 9
	AUC	0.760 0	0.670 0	0.810 0	0.840 0	0.930 0	0.950 0	0.950 0
KNN	G-mean	0.467 6	0.870 0	0.895 2	0.903 5	0.917 4	0.918 9	0.913 2
	AUC	0.570 0	0.870 0	0.880 0	0.900 0	0.940 0	0.940 0	0.940 0
SVM	G-mean	0.000 0	0.726 3	0.767 9	0.792 1	0.939 1	0.939 1	0.939 1
	AUC	0.580 0	0.910 0	0.810 0	0.840 0	0.930 0	0.910 0	0.920 0
NN	G-mean	0.486 0	0.869 8	0.886 7	0.899 7	0.876 6	0.864 6	0.873 4
	AUC	0.6400	0.910 0	0.910 0	0.910 0	0.920 0	0.920 0	0.920 0

表4 NaN-Bicluster SMOTE与对比模型的实验结果(k=7)

Table 4 Experimental results of NaN-Bicluster SMOTE and contrast models (k=7)

模型	评价准则	无采样模型参与	SMOTE	Borderline SMOTE	Safe-Level SMOTE	NaN-Bicluster SMOTE		
						$\delta=0.01$	$\delta=0.10$	$\delta=0.30$
DT	G-mean	0.000 0	0.897 9	0.856 3	0.870 7	0.920 6	0.927 2	0.926 2
	AUC	0.500 0	0.930 0	0.890 0	0.890 0	0.960 0	0.960 0	0.950 0
LR	G-mean	0.501 0	0.503 0	0.758 5	0.767 1	0.907 3	0.930 1	0.930 2
	AUC	0.760 0	0.650 0	0.820 0	0.840 0	0.940 0	0.960 0	0.960 0
KNN	G-mean	0.467 6	0.893 0	0.898 4	0.901 1	0.920 1	0.922 4	0.921 9
	AUC	0.570 0	0.890 0	0.880 0	0.890 0	0.950 0	0.950 0	0.950 0
SVM	G-mean	0.000 0	0.645 2	0.768 0	0.784 5	0.939 1	0.939 1	0.939 1
	AUC	0.580 0	0.740 0	0.810 0	0.840 0	0.940 0	0.930 0	0.930 0
NN	G-mean	0.486 0	0.892 6	0.894 1	0.903 0	0.892 1	0.888 9	0.895 0
	AUC	0.640 0	0.920 0	0.910 0	0.910 0	0.940 0	0.940 0	0.940 0

部规则改进合成公式的方式能有效提升均衡能力。

通过表3的实验结果可以看出，NaN-Bicluster SMOTE的均衡能力仅在G-mean评价准则上，劣于Safe-Level SMOTE+NN。相较于$k=3$，NaN-Bicluster SMOTE的均衡能力得到较大幅度的提升。这说明随着均衡比例k的提升，机器学习模型对于NaN-Bicluster SMOTE均衡后的数据集具有更强的分类性能。

从表4中可以看出，NaN-Bicluster SMOTE的总体均衡能力相较于其余模型而言，有着明显优势，且随着k的增加，不会出现均衡能力的弱化。再次说明了考虑合成范围设定、以局部规则改进合成公式的方式能有效提升均衡能力，且在均衡比例$k=7$时尤为突显。

以OBT-Adaptive-SVM为对比模型，该模型需规定SVM核函数为RBF、固定删除系数α为0.3，对SVM惩罚因子C及RBF超参数γ进行寻优，具体参数取值如表5所示。其中，NaN-Bicluster SMOTE的实验结果来自表2～4中与SVM结合时的最大值。

表5　NaN-Bicluster SMOTE与对比模型OBT-Adaptive-SVM的实验结果

Table 5　Experimental results of NaN-Bicluster SMOTE and OBT-Adaptive-SVM

均衡比例	OBT-Adaptive-SVM						NaN-Bicluster SMOTE-SVM	
	核函数	C	γ	α	G-mean	AUC	G-mean	AUC
$k=3$	RBF	1 000	1.0	0.3	0.895 2	0.910 0	0.939 1	0.890 0
$k=5$	RBF	1 000	2.0	0.3	0.913 8	0.920 0	0.939 1	0.930 0
$k=7$	RBF	1 000	2.0	0.3	0.917 3	0.920 0	0.939 1	0.940 0

通过观察表5可知,结合过采样和欠采样、融合自适应变核参数的OBT-Adaptive-SVM,其均衡能力相较于SMOTE、Borderline SMOTE、Safe-Level SMOTE具有明显提升。更需要注意的是,OBT-Adaptive-SVM通过参数寻优实现性能的最大化,但最优参数往往过大,易出现过拟合情况,性能、可解释性和现实层面的可操作性弱于本文提出的NaN-Bicluster SMOTE。

根据上述实验结果总结出如下结论:

(1) 基于NaN-Bicluster SMOTE的模型性能总体上优于对比模型。在不同的参数k和δ的情况下,基于NaN-Bicluster SMOTE的机器学习模型的G-mean、AUC相较于对比模型,均有较大幅度的提升,说明了本文所构建的NaN-Bicluster SMOTE方法的有效性。

(2) NaN-Bicluster SMOTE适用于不同的机器学习模型且性能优越。无论与结构单一的DT、LR和KNN结合,还是与结构多样、计算复杂的SVM、NN结合,NaN-Bicluster SMOTE均适用且发挥了较好的性能。

(3) NaN-Bicluster SMOTE的均衡能力不会随着k增加,出现退化的情况。随着k逐渐增加至正负类样本均衡的状态时,合成样本的数量出现急剧的增长,过多合成样本的加入使样本间相似性增强、区分性减弱,从而存在干扰机器学习模型的分类性能。例如,基于Safe-Level SMOTE的LR、SVM机器学习模型,其G-mean、AUC随着k的增加出现明显的下降趋势。而基于NaN-Bicluster SMOTE的机器学习模型规避了过多合成样本造成数据混沌的弊端,其G-mean、AUC保持稳定的趋势。

(4) NaN-Bicluster SMOTE的局部规则提取中的最大平均平方残差δ虽然会使NaN-Bicluster SMOTE的均衡能力出现波动,但该波动幅度是轻微的,使得不同δ的NaN-Bicluster SMOTE的均衡能力处于稳定水平,且优于其余对比模型。

(5) NaN-Bicluster SMOTE适用于非均衡的信贷数据,能有效提升机器学习模型在该类数据集上的分类性能。通过自然近邻选取参与采样的样本与设定安全范围,使用双聚类提取局部规则以此改进SMOTE合成公式的方式,提升了SMOTE的均衡能力,拓宽了SMOTE的应用场景。

3.4　模型统计检验

为进一步检验NaN-Bicluster SMOTE是否与对比模型存在显著差异,是否具有较强的均衡能力。选用统计检验方法中的非参数Wilcoxon显著性检验,将NaN-Bicluster SMOTE与对比模型进行两两配对检验。鉴于OBT-Adaptive-SVM结果较少,不足以进行该检验,所以将SMOTE、Borderline SMOTE和Safe-Level SMOTE的具体检验结果统计列于表6中。

表6　Wilcoxon显著性检验结果

Table 6　Results of Wilcoxon significance test

δ	评价准则	SMOTE	Borderline SMOTE	Safe-Level SMOTE
0.01	G-mean	0.001	0.005	0.009
	AUC	0.001	0.002	0.012
0.10	G-mean	0.001	0.005	0.011
	AUC	0.001	0.002	0.010
0.30	G-mean	0.001	0.004	0.011
	AUC	0.001	0.002	0.010

从表6中可以看出,所有Wilcoxon显著性检

验的 P 值均低于 0.05，说明 NaN-Bicluster SMOTE 与 SMOTE、Borderline SMOTE 和 Safe-Level SMOTE存在显著性的差异，进一步验证了NaN-Bicluster SMOTE的均衡能力较优且与对比模型存在明显的区别。

4 结束语

随着小额贷款行业的蓬勃发展，借助非均衡的信贷数据，有效甄别“违约”借款人已成为信用风险评估研究中的重中之重，有利于保障决策者和投资者的利益，促进行业的健康长远发展。考虑到SMOTE处理非均衡数据的不足，本文基于自然近邻构建采样样本选取、合成范围设定的逻辑规则，解决了现有研究中依赖参数设定的弊端和填补了安全范围设定的空白；双聚类提取局部规则以改进合成公式，使合成样本继承局部规则，实现了合成公式可解释性和完善性的大幅度提升；在Prosper信贷数据集上进行多维度对比试验并验证NaN-Bicluster SMOTE的均衡性能，扩展了模型在现实中的应用范畴。

对于信贷行业决策者而言，NaN-Bicluster SMOTE可应用于借贷行为之前，搜集借款人必要信息以初步评价其信用风险，结合实地调研的人工评价，进行综合性决断。NaN-Bicluster SMOTE也反向要求信贷行业决策者重点关注“违约”的少数类借款人；要求企业建立体系化服务模式，做好贷前收集信息、贷中更新信息、贷后再训练等系统性工作，从而提升决策的精准性、可靠性和运营效率。

此外考虑到决策者从众多参数组合下的NaN-Bicluster SMOTE选择合适的模型需要花费大量的时间、精力的局限性，未来工作将进一步研究动态自适应参数的NaN-Bicluster SMOTE方法，以辅助决策者进行有效甄别，规避潜在损失。同时，为了便于决策者在实际中的应用，封装式、一体化模型也是未来的研究重点之一。

参考文献：

[1] 杜晓山，孙若梅. 中国小额信贷的实践和政策思考[J]. 财贸经济，2000(7): 32-37.
DU Xiaoshan, SUN Ruomei. Practice and policy thinking of microfinance in China[J]. Finance and Trade Economics, 2000(7): 32-37.

[2] 张博，胡金焱，马驰骋. 从钱庄到小额贷款公司:中国民间金融发展的历史持续性[J]. 经济学，2018，17(4): 1383-1408.
ZHANG Bo, HU Jinyan, MA Chicheng. Where traditional banks meet modern micro-credit: The long-term persistence of informal finance in China[J]. China Economic Quarterly, 2018, 17(4): 1383-1408.

[3] 胡贤德，曹蓉，李敬明，等. 小微企业信用风险评估的IDGSO-BP集成模型构建研究[J]. 运筹与管理，2017，26(4): 132-139，148.
HU Xiande, CAO Rong, LI Jingming, et al. Research into the credit evaluation model of small and micro businesses based on IDGSO-BP comprehensive method[J]. Operations Research and Management Science, 2017, 26(4): 132-139, 148.

[4] 石宝峰，刘锋，王建军，等. 基于PROMETHEE-Ⅱ的商户小额贷款信用评级模型及实证[J]. 运筹与管理，2017，26(9): 137-147.
SHI Baofeng, LIU Feng, WANG Jianjun, et al. A credit rating model of microfinance loans for small private business based on PROMETHEE-Ⅱ and its empirical study[J]. Operations Research and Management Science, 2017, 26(9): 137-147.

[5] 陈良维. 决策树算法在农户小额贷款中的应用研究[J]. 计算机工程与应用，2008(31): 242-244，248.
CHEN Liangwei. Study and development of decision-tree algorithm on farmer credit evaluation[J]. Computer Engineering and Applications, 2008(31): 242-244, 248.

[6] 迟国泰，龚玲玲. 商户小额贷款决策模型[J]. 技术经济，2016，35(4): 98-103.
CHI Guotai, GONG Lingling. Decision model on petty loan for merchants[J]. Journal of Technology Economics, 2016, 35(4): 98-103.

[7] LI B, ZHANG X, YAN J, et al. Application of KNN algorithm based on value difference metric and clustering optimization in

bank customer behavior prediction[J]. Journal of Computer Applications, 2019, 39(9): 2784-2788.

[8] 衣柏衡, 朱建军, 李杰. 基于改进SMOTE的小额贷款公司客户信用风险非均衡SVM分类[J]. 中国管理科学, 2016, 24(3): 24-30.
YI Baiheng, ZHU Jianjun, LI Jie. Imbalanced data classification on micro-credit company customer credit risk assessment using improved SMOTE support vector machine[J]. Chinese Journal of Management Science, 2016, 24(3): 24-30.

[9] 吴冲, 郭英见, 夏晗. 基于模糊积分支持向量机集成的商业银行信用风险评估模型研究[J]. 运筹与管理, 2009, 18(2): 115-119.
WU Chong, GUO Yingjian, XIA Han. The model of credit risk assessment in commercial banks on fuzzy integral support vector machines ensemble[J]. Operations Research and Management Science, 2009, 18(2): 115-119.

[10] BAESENS B, SETIONO R, MUES C, et al. Using neural network rule extraction and decision tables for credit-risk evaluation[J]. Management Science, 2003, 49(3): 312-329.

[11] TAO X, LI Q, GUO W, et al. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification[J]. Information Sciences, 2019, 487: 31-56.

[12] ZHENG M, LI T, SUN L, et al. An automatic sampling ratio detection method based on genetic algorithm for imbalanced data classification[J]. Knowledge-Based Systems, 2021, 216: 106800.

[13] PAN T, ZHAO J, WU W, et al. Learning imbalanced datasets based on SMOTE and Gaussian distribution[J]. Information Sciences, 2020, 512: 1214-1233.

[14] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[15] SUN J, LANG J, FUJITA H, et al. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates[J]. Information Sciences, 2018, 425: 76-91.

[16] 缪裕青, 欧威健, 刘同来, 等. 基于情感极性与SMOTE过采样的虚假评论识别方法[J]. 计算机应用研究, 2018, 35(7): 2042-2045.
MIAO Yuqing, OU Weijian, LIU Tonglai, et al. Detection of fake reviews based on sentiment polarity and over-sampling[J]. Application Research of Computers, 2018, 35(7): 2042-2045.

[17] PARSA A B, MOVAHEDI A, TAGHIPOUR H, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis[J]. Accident Analysis and Prevention, 2019, 136: 105405.

[18] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences, 2019, 501: 118-135.

[19] MAJZOUB H A, ELGEDAWY I, YKU A, et al. HCAB-SMOTE: A hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification[J]. Arabian Journal for Science and Engineering, 2020, 45(4): 3205-3222.

[20] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]// Proceedings of International Conference on Intelligent Computing. Berlin, Germany: Springer, 2005: 878-887.

[21] BUNKHUMPORNPAT C, SINAPIROMSARAN K, LURSINSAP C. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem[C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Germany: Springer, 2009: 475-482.

[22] 赵冬雪, 王昕, 王利东. 基于属性选择和采样策略的不平衡数据动态分类方法[J]. 数据采集与处理, 2021, 36(3): 509-518.
ZHAO Dongxue, WANG Xin, WANG Lidong. Dynamic classification for multi-imbalanced datasets via attribute selection and sampling strategy[J]. Journal of Data Acquisition and Processing, 2021, 36(3): 509-518.

[23] 邢延, 陈嘉锋, 贾小彦, 等. 类别混叠度对非均衡数据分类的有效性分析[J]. 数据采集与处理, 2018, 33(5): 936-944.
XING Yan, CHEN Jiafeng, JIA Xiaoyan, et al. Evaluation of class overlap measures on imbalanced data classification[J]. Journal of Data Acquisition and Processing, 2018, 33(5): 936-944.

[24] 吴志峰, 黄若尘, 魏昕, 等. 非均衡IPTV数据集下的用户报障预测[J]. 数据采集与处理, 2018, 33(1): 75-84.
WU Zhifeng, HUANG Ruochen, WEI Xin, et al. Prediction for user's complaint in imbalanced IPTV dataset[J]. Journal of Data Acquisition and Processing, 2018, 33(1): 75-84.

[25] LI J, ZHU Q, WU Q, et al. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural

neighbors[J]. Information Sciences, 2021, 565: 438-455.

[26] SOLTANZADEH P, HASHEMZADEH M. RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem[J]. Information Sciences, 2021, 542: 92-111.

[27] PADILHA V A, CAMPELLO R J G B. A systematic comparative evaluation of biclustering techniques[J]. BMC Bioinformatics, 2017, 18(1): 1-25.

[28] 黄学文, 孙榕, 艾亚晴. 招标采购中的采购物品打包模型及其优化算法[J]. 运筹与管理, 2020, 29(9): 18-26.
HUANG Xuewen, SUN Rong, AI Yaqing. Modelling and optimization algorithm for bundling from buyer's perspective in procurement[J]. Operations Research and Management Science, 2020, 29(9): 18-26.

[29] ZHU Q, FENG J, HUANG J. Natural neighbor: A self-adaptive neighborhood method without parameter K[J]. Pattern Recognition Letters, 2016, 80: 30-36.

[30] HUANG J, ZHU Q, YANG L, et al. A non-parameter outlier detection algorithm based on natural neighbor[J]. Knowledge-Based Systems, 2016, 92: 71-77.

[31] ZHANG Y, DING S, WANG Y, et al. Chameleon algorithm based on improved natural neighbor graph generating sub-clusters[J]. Applied Intelligence, 2021, 51(11): 8399-8415.

[32] LI J, ZHU Q, WU Q, et al. SMOTE-NaN-DE: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution[J]. Knowledge-Based Systems, 2021, 223: 107056.

[33] 崔衍, 薛源. 基于改进蝙蝠算法的双聚类算法设计与实现[J]. 现代计算机, 2021, 27(30): 38-44.
CUI Yan, XUE Yuan. Design and implementation of bi-clustering algorithm based on improved bat algorithm[J]. Modern Computer, 2021, 27(30): 38-44.

[34] HUANG Q, HU B, ZHANG F. Evolutionary optimized fuzzy reasoning with mined diagnostic patterns for classification of breast tumors in ultrasound[J]. Information Sciences, 2019, 502: 525-536.

[35] 王星, 王峻, 余国先, 等. 基于网络约束双聚类的癌症亚型分类[J]. 计算机学报, 2019, 42(6): 1274-1288.
WANG Xing, WANG Jun, YU Guoxian, et al. Network regularized bi-clustering for cancer subtype categorization[J]. Chinese Journal of Computers, 2019, 42(6): 1274-1288.

[36] DORE T, MACH T. Marketplace lending and consumer credit outcomes: Evidence from prosper[J]. Applied Economics, 2022, 54(4): 390-405.

[37] GUO H, LIU H, WU C, et al. Logistic discrimination based on G-mean and F-measure for imbalanced problem[J]. Journal of Intelligent and Fuzzy Systems, 2016, 31(3): 1155-1166.

[38] RI J H, KIM H. G-mean based extreme learning machine for imbalance learning[J]. Digital Signal Processing, 2019, 98: 102637.

[39] NORTON M, URYASEV S. Maximization of AUC and buffered AUC in binary classification[J]. Mathematical Programming, 2019, 174(1): 575-612.

作者简介:

何亮(1997-),男,硕士研究生,研究方向:数据分析、机器学习,E-mail: heliang@nuaa.edu.cn。

徐海燕(1963-),**通信作者**,女,教授,研究方向:数据分析、冲突分析,E-mail: xuhaiyan@nuaa.edu.cn。

陈璐(1994-),女,博士研究生,研究方向:冲突分析、模糊决策。

(编辑:王静)