计算机科学
COMPUTER SCIENCE

基于方差迁移的非平衡数据过采样方法

郑一凡, 王卯宁

引用本文

郑一凡, 王卯宁. 基于方差迁移的非平衡数据过采样方法[J]. 计算机科学, 2024, 51(6A): 230400198-6.

ZHENG Yifan, WANG Maoning. Imbalanced Data Oversampling Method Based on Variance Transfer[J]. Computer Science, 2024, 51(6A): 230400198-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400198

基于方差迁移的非平衡数据过采样方法

郑一凡　王卯宁
中央财经大学信息学院　北京 102206
(zhengyf_cufe@163.com)

摘　要　重采样是解决非平衡数据分类问题的重要方法。但在数据集很小的情况下,欠采样会丢失数据集的重要信息,因此过采样是非平衡数据分类问题的研究重点。现有的过采样方法虽然部分解决了类间不平衡问题,但是本质上并未给少数类引入额外的信息,且仍然存在着过拟合的风险。针对这些问题,提出了一种基于多数类方差迁移的少数类合成方法(Variance Transfer Oversampling,VTO),从足够多样化的多数类中提取样本偏移向量,综合少数类和多数类的特征权重矩阵以调整,最终将经过置信条件筛选的偏移向量叠加至少数类样本中心,从而在少数类样本生成中引入多数类方差,进而丰富少数类特征空间。为了验证所提算法的有效性,使用决策树为分类模型在6个KEEL数据集上训练,对比SMOTEENN等其他过采样方法,以F-score和PR-AUC值为评价指标进行了实验。结果显示,该算法在处理非平衡数据分类问题时具有更大优势。

关键词:非平衡数据;分类;过采样;方差迁移;协方差
中图分类号　TP311

Imbalanced Data Oversampling Method Based on Variance Transfer

ZHENG Yifan and WANG Maoning
School of Information,Central University of Finance and Economics,Beijing 102206,China

Abstract　Resampling is an important method to solve imbalanced data classification problem. However,when the size of data set is very small,undersampling will lose important information of the data set,so oversampling is the research focus of imbalanced data classification. Although the existing oversampling methods partially solve the problem of imbalance between classes,they essentially do not introduce additional information to minority class,and there is still a risk of overfitting. To solve these problems,VTO,an oversampling method based on variance migration of the majority class,is proposed in this paper. In this method,a shift vector is extracted from majority class,and the feature weight matrix of the minority class and the majority class is used for adjustment. Furthermore,the shift vectors filtered by the confidence conditions are superimposed to the center of the minority class,so as to introduce the majority class variance in the generation process of new minority class samples,then enrich the minority class feature space. In order to verify the effectiveness of the proposed algorithm,decision tree is used as classification model to train on 6 KEEL data sets. Compared with SMOTEENN and other over-sampling methods,with F-score and PR-AUC values as evaluation indexes,the results show that VTO is more advantageous in dealing with imbalanced data classification.

Keywords　Imbalanced data,Classification,Oversampling,Variance transfer,Covariance

1　引言

任何在多数类和少数类之间分布不均的数据集都可以被认为存在类不平衡,多数类占据了数据集的大部分,少数类只有有限的数据表示,但它们往往是研究者更感兴趣的类。真实世界的数据常常存在类别不平衡现象,典型场景如欺诈检测[1]、故障检测[2]、医疗识别[3]等。在这些领域中,那些需要重点识别的数据只占总数量极小的一部分。由于数据分布不平衡,大多数情况下,少数类样本的不充足导致其内含信息的丰富度不够,在分类时很难正确分析出数据的分布及其内部规律,使模型的分类效果受到影响。因此,研究非平衡数据集分类问题以实现少数类样本的高效分类,对提高部分现实场景中的重点类的识别有重要意义。

目前,对非平衡数据的处理方法大致分为两类,即基于数据处理的方法和基于算法改进的方法。基于数据处理的方法主要使用重采样技术将样本数量重平衡,包括:(1)欠采样:丢弃部分多数类样本,使其减少到少数类的样本数量;(2)过采样:在少数类中加入新的样本,使其和多数类的样本数量达到平衡;(3)混合采样:既进行少数类样本的增加,又进行多数类样本的消减。基于算法改进的方法目的是使原本适用于平衡数据的算法侧重对少数类的学习,主要有:(1)代价敏感:将错分少数类赋予比错分多数类更大的代价,以使算法更倾向于

基金项目:国家自然科学基金(61907042,61702570);北京市自然科学基金(4194090);四川省教育厅人文社会科学重点研究基地科技金融与创业金融研究中心课题(JR2018-2))
This work was supported by the National Natural Science Foundation of China(61907042,61702570),Beijing Natural Science Foundation(4194090) and Project of Research Center for Science and Technology Finance and Entrepreneurship Finance,Key Research Base of Humanities and Social Sciences,Sichuan Provincial Department of Education(JR2018-2).
通信作者:王卯宁(13854139297@139.com)

提高少数类的分类准确率;(2)集成学习:通过结合多个模型的预测结果来提高分类性能[4]。

由于重采样技术简单、直观,而且改变的是数据集本身,而不是分类器,因此大多数研究者在解决非平衡数据的分类问题时,首先会使用重采样进行初步的处理,但目前的方法都存在着可改进的空间。欠采样的典型算法,如基于聚类和图技术处理的 DBIG-US[5],基于质心空间的不均衡数据欠采样方法 ICIKMDS[6],在执行时选择性丢弃了数据集的部分样本,这种方法适用于数据量充足的数据集,但在小数据集场景的应用中进行欠采样将损失珍贵的非欺诈样本信息,进而极大地影响模型的分类效果,降低分类模型的有效性。

相比之下,过采样技术不会丢失数据集信息[7],更受到小数据集研究者的青睐。因此,本文主要对适用于小样本情形的过采样方法进行探索研究。在每个少数类样本与其近邻样本的连线上合成新样本的经典过采样算法 SMOTE[8],与在其基础上的改进方法形成 SMOTE 族,包括:根据其与最近多数类样本距离分配权重,利用聚类合成位于少数类内新样本的 MWMOTE[9];旨在通过添加局部离群因子(Local Outlier Factor)来提升对合成样本中噪声识别能力的 SMOTE-LOF[10];保留含有大量少数样本的聚类以合成样本,然后放入少数类样本不足的聚类中的 Kmeans SMOTE[11]等。除此之外,文献[12]针对少数类密集层的边界和稀疏层进行双向过采样,也在一定程度上提高了分类性能。

然后,以上这些过采样算法均基于少数类内信息来生成新样本,并未引入额外的信息,实际上并未增加少数类样本的丰富度和信息量。虽然总体性能得到了改善,但它们无法从根本上解决数据量有限的少数类中信息缺少的问题。针对该限制的一个可行的解决方案是对少数类进行信息增强,即在模型训练中引入额外的信息,以便在不平衡分类中提高模型性能。迁移学习(Transfer Learning)试图从源域(例如数据集、任务或类)向目标域迁移知识,以增强目标域上的模型训练。小样本学习为迁移学习思想的一个典型研究与应用方向[13]。因此,为了更好地解决小样本数据集中存在的不平衡问题,可以考虑从多数类到少数类的知识迁移方法,以对小样本类别不平衡问题做出针对性处理。

2 一种基于方差迁移的过采样方法

2.1 算法思想

针对以上问题和思考,本文提出了一种基于多数类方差迁移的过采样方法(Variance Transfer Oversampling,VTO),旨在进行从多数类到少数类的信息迁移,丰富少数类内方差,以增强少数类上的模型性能。在大部分非平衡数据集中,少数类通常由于样本量的限制,并不能展现该类的真实分布,具有较低的类内方差,而这种不均匀的分布扭曲了整体的特征空间[14]。当类别不平衡存在时,特征分布与类样本数密切相关,由于少数类样本稀少,不能为模型的学习提供足够的类内多样性,因此会降低模型对少数类的分类效果;而多数类具有充足的样本,在特征空间的跨度更大,贴合真实分布,具有更大的类内方差,因此多数类可以被更好地区分[14]。故而,本文考虑迁移从多数类获取到的类内方差,即在合成少数类新样本时向少数类中引入多数类的方差信息,以丰富少数类特征空间,从而提升分类效果。该算法的描述如算法 1 所示。

算法 1 基于方差迁移的过采样算法 VTO

输入:多数类特征向量集合 $D_{major}=\{\alpha_1,\alpha_2,\cdots,\alpha_{n_{major}}\}$,少数类特征向量集合 $D_{minor}=\{\beta_1,\beta_2,\cdots,\beta_{n_{major}}\}$,主成分个数 φ,参数 K,置信概率阈值 ρ

输出:新生成少数类样本 $D_{new}=\{\gamma_1,\gamma_2,\cdots,\gamma_N\}$

1. $N=n_{major}-n_{minor}$//计算应生成的少数样本数量
2. $\mu_{major}=ave(D_{major})$ //计算多数类样本中心点
3. $\mu_{minor}=ave(D_{minor})$ //计算少数类样本中心点
4. for α_i in D_{major}//计算多数类各样本对样本中心的偏移向量
5. $\quad\Delta_i=\|\alpha_i-\mu_{major}\|$
6. end for
7. for i=1,2,…,n_{major}
8. $\quad V_1+=(\alpha_i-\mu_{major})^T(\alpha_i-\mu_{major})$ //计算多数类内协方差矩阵
9. end for
10. $Q_1=PCA(V_1,\varphi)$ //计算多数类因子载荷矩阵
11. 同理计算少数类因子载荷矩阵 Q_2
12. for α_i in D_{major}//计算新偏移向量
13. $\quad\widetilde{\Delta}_i=K\times\Delta_i Q_1 Q_1^T+(1-K)\times\Delta_i Q_2 Q_2^T$
14. $\quad$A. append(满足筛选规则 1 和规则 2 的 $\widetilde{\Delta}_i$)
15. end for
16. for $\widetilde{\Delta}_i$ in random(A,N) //随机抽取 A 中 N 个偏移向量做方差迁移
17. $\quad\gamma_i=\mu_{minor}+\widetilde{\Delta}_i$
18. end for
19. end

2.2 算法流程

首先,定义数据集 $T=\{(\alpha_1,y_1),(\alpha_2,y_1),\cdots,(\alpha_{n_{major}},y_1),(\beta_1,y_2),(\beta_2,y_2),\cdots,(\beta_{n_{major}},y_2)\}$,其中 $\alpha_i,\beta_j\in\mathbb{R}^d$,分别为少数类与多数类的特征向量;$y_k\in\{0,1\}$ 为类别标签;$i=1,2,\cdots,n_{major}$,$j=1,2,\cdots,n_{minor}$,且 $n_{major}>n_{minor}$,即多数类样本数量大于少数类样本数量。多数类 $D_{major}=\{\alpha_1,\alpha_2,\cdots,\alpha_{n_{major}}\}$,$\alpha_i\in D_{major}$;少数类 $D_{minor}=\{\beta_1,\beta_2,\cdots,\beta_{n_{major}}\}$,$\beta_j\in D_{minor}$;单个样本中的特征数为 n_f,即 $\alpha_i=[a_1\ \ a_2\ \ \cdots\ \ a_{n_f}]$,$\beta_j=[b_1\ \ b_2\ \ \cdots\ \ b_{n_f}]$。

基于中心极限定理,不妨假设多数类与少数类均服从多元高斯分布,即 $\alpha\sim N(\mu_{major},\Sigma_{major})$,$\beta\sim N(\mu_{minor},\Sigma_{minor})$,各特征维度 $f_i\sim N(u_i,\sigma_i^2)$。算法的目标是将多数类方差迁移至少数类,即创建新少数类样本 $\gamma\sim N(\mu_{minor},\Sigma_{major})$,如图 1 所示。具体有以下两种迁移策略,如图 2 所示。

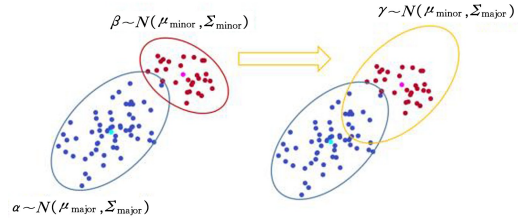


图 1 方差迁移目标

Fig. 1 Variance transfer target

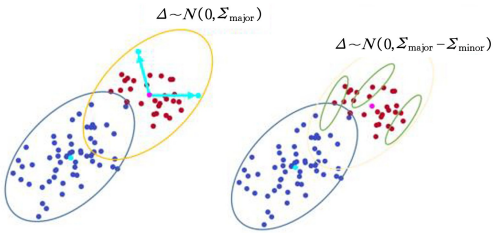


图 2 两种迁移策略示意图

Fig. 2 Two different variance transfer strategies

迁移策略 1　以μ_{minor}为中心,叠加服从$N(0,\Sigma_{\text{major}})$的偏移向量;

迁移策略 2　以β_j为中心,叠加服从$N(0,\Sigma_{\text{major}}-\Sigma_{\text{minor}})$的偏移向量。

本文采用策略 1 来进行多数类到少数类的方差迁移,以使新生成样本满足多数类的方差分布;同时,利用策略 2 来划定新样本生成置信区间,即仅将落在以某一样本为中心的$N(0,\Sigma_{\text{major}}-\Sigma_{\text{minor}})$置信区间内的新增样本予以保留。具体迁移步骤如下:

步骤 1　计算各类中心。将总样本计算平均值,分别得到多数类和少数类的类中心点μ_{major},μ_{minor},定义如下。

定义 1　$\forall \alpha_i \in D_{\text{major}}$,根据多数类样本特征向量算术平均值计算,则多数类类中心点μ_{major}可以表示为:

$$\mu_{\text{major}}=\frac{1}{|D_{\text{major}}|}\sum_{i=1}^{n_{\text{major}}}\alpha_i,\alpha_i\in D_{\text{major}} \tag{1}$$

同理,少数类类中心点μ_{minor}可表示为:

$$\mu_{\text{minor}}=\frac{1}{|D_{\text{minor}}|}\sum_{j=1}^{n_{\text{major}}}\beta_j,\beta_j\in D_{\text{minor}} \tag{2}$$

步骤 2　计算用于迁移的多数类样本偏移向量。初始偏移向量定义如下。

定义 2　$\forall \alpha_i \in D_{\text{major}}$,其初始偏移向量为:

$$\Delta_i=\alpha_i-\mu_{\text{major}},\alpha_i\in D_{\text{major}} \tag{3}$$

其中,Δ_i为服从$N(0,\Sigma_{\text{major}})$分布的随机向量。

为提取类内各特征重要性,本文计算各类样本不同特征之间的协方差,得到类内协方差矩阵。多数类协方差矩阵计算如下:

$$\begin{aligned}V_1&=\sum_{i=1}^{n_{\text{major}}}(\alpha_i-\mu_{\text{major}})^{\mathrm{T}}(\alpha_i-\mu_{\text{major}})\\&=\sum_{i=1}^{n_{\text{major}}}\begin{bmatrix}a_{i1}\\a_{i2}\\\vdots\\a_{in_f}\end{bmatrix}\begin{bmatrix}a_{i1}&a_{i2}&\cdots&a_{in_f}\end{bmatrix}\\&=\sum_{i=1}^{n_{\text{major}}}\begin{bmatrix}a_{i1}^2&a_{i1}a_{i2}&\cdots&a_{i1}a_{in_f}\\a_{i1}a_{i2}&a_{i2}^2&\cdots&a_{i2}a_{in_f}\\\vdots&\vdots&\ddots&\vdots\\a_{i1}a_{in_f}&a_{i2}a_{in_f}&\cdots&a_{in_f}^2\end{bmatrix}\end{aligned} \tag{4}$$

同理,得到少数类类内协方差矩阵V_2。

随后,对多数类协方差矩阵进行主成分分析,以使新生成样本在差异性更大的特征上更突出。根据需提取主成分个数φ,计算因子载荷矩阵Q_1:

$$Q_1=PCA(V,\varphi),\varphi\in\{1,2,\cdots,n_f\} \tag{5}$$

同理得Q_2。因子载荷矩阵为正交矩阵,其每一列对应一个主成分,数值表示了原始特征在相应主成分上的贡献程度。

随后,本文根据式(15)计算用于迁移的新偏移向量。QQ^{T}为对称矩阵,Δ_iQQ^{T}可将Δ_i中各特征分量根据该类特征重要性进行调整;当$\varphi=n_f$,即未对协方差矩阵V做降维处理时,QQ^{T}为单位矩阵,新偏移向量$\widetilde{\Delta}_i=\Delta_i$;参数$K$决定调整时依据少数类或多数类的特征重要性程度。

$$\begin{aligned}\widetilde{\Delta}_i&=K\times\Delta_iQ_1Q_1^{\mathrm{T}}+(1-K)\times\Delta_iQ_2Q_2^{\mathrm{T}}\\&=\begin{bmatrix}c_{i1}&c_{i2}&\cdots&c_{ij}&\cdots&c_{in_f}\end{bmatrix}\end{aligned} \tag{6}$$

步骤 3　过滤迁移向量。为了控制生成少数类样本的方差真实性,扩大少数类特征空间并排除离群点对分布的影响,需对用于迁移的偏移向量$\widetilde{\Delta}_i$进行筛选。对于多元高斯分布中的各个特征分量,在多数类中$f_{1t}\sim N(u_{1t},\sigma_{1t}^2)$,在少数类中$f_{2t}\sim N(u_{2t},\sigma_{2t}^2)$,$t\in\{1,2,\cdots,n_f\}$。在大部分特征上,多数类由于其样本量丰富,类内方差大于少数类,应进行方差迁移;少部分特征上,多数类类内方差小于少数类,因此无需进行迁移。据此,本文设定如下筛选规则。

筛选规则 1　对于特征f_t,$\widetilde{\Delta}_i$在该特征上的分量为c_{it};若多数类方差,大于少数类方差即$\sigma_{1t}^2>\sigma_{2t}^2$,$c_{it}$需在以任一真实少数类样本$\beta_j$在该特征上的分量$b_{jt}$为中心的$N(b_{jt},\sigma_{1t}^2-\sigma_{2t}^2)$置信区间内,即代入概率密度函数积分计算结果满足:

$$2\times\int_{c_{it}}^{+\infty}\frac{1}{\sqrt{2\pi}\sqrt{\sigma_{1t}^2-\sigma_{2t}^2}}\exp\left(-\frac{(x-b_{jt}-\mu_{\text{minort}})^2}{2(\sigma_{1t}^2-\sigma_{2t}^2)}\right)\mathrm{d}x<\rho \tag{7}$$

筛选规则 2　对于特征f_j,若多数类方差小于等于少数类方差,即$\sigma_{1t}^2\leqslant\sigma_{2t}^2$,则无需丰富少数类在该特征上的方差,$c_{ij}$只需在$N(0,\sigma_{2t}^2)$置信区间内,即代入概率密度函数积分计算结果满足:

$$2\times\int_{c_{it}}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{2t}}\exp\left(-\frac{x^2}{2\sigma_{2t}^2}\right)\mathrm{d}x<\rho \tag{8}$$

在具体代码实现中,首先计算出给定置信概率阈值$\rho\in[0,1]$下的标准正态分位数z,计算结果如表 1 所列。图 3 为标准正态分布示意图。

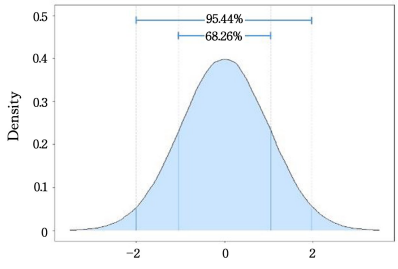


图 3　标准正态分布置信度与相应置信区间

Fig. 3　Standard normal distribution confidence and corresponding confidence interval

表 1　置信度与置信区间示例

Table1　Examples of confidence and confidence intervals

置信度ρ/%	标准正态分位数z	标准置信区间
38	0.5	$[\mu-0.5\sigma,\mu+0.5\sigma]$
68	1	$[\mu-\sigma,\mu+\sigma]$
86	1.5	$[\mu-1.5\sigma,\mu-1.5\sigma]$
95	2	$[\mu-2\sigma,\mu-2\sigma]$

随后,对于数据集的每个特征维度,如在该特征上多数类方差大于少数类方差,则适用筛选规则 1,每一个$\widetilde{\Delta}_i$在该特征上的分量c_{ij}需在以任一真实少数类样本β_j为中心的$N(b_{jt},\sigma_{1t}^2-\sigma_{2t}^2)$置信区间内。逆向考虑,可判断是否存在至少一个少数类样本β_j,满足其在每一个多数类方差大于少数类方差的特征上的分量均落入$N(c_{it},\sigma_{1t}^2-\sigma_{2t}^2)$置信区间$[c_{it}-z*\sqrt{\sigma_{1t}^2-\sigma_{2t}^2},c_{it}+z*\sqrt{\sigma_{1t}^2-\sigma_{2t}^2}]$内。

如在该特征上多数类方差小于等于少数类方差,则直接适用筛选规则 2,判断$\widetilde{\Delta}_i$在该特征上的分量c_{it}在$N(0,\sigma_{2t}^2)$置信区间$[-z\sigma,+z\sigma]$内。

若有$\widetilde{\Delta}_i$满足以上条件,则将该$\widetilde{\Delta}_i$纳入最终迁移集合A。

步骤 4　执行迁移,生成新样本。计算需生成的少数类

样本数量，即原始多数类样本数量与少数类样本数量之差：

$$N = n_{\text{major}} - n_{\text{minor}} \tag{9}$$

随机抽取最终迁移集合 A 中的 N 个偏移向量 $\widetilde{\Delta}_i \in A$，进行方差迁移：

$$\gamma_i = \mu_{\text{minor}} + \widetilde{\Delta}_i \tag{10}$$

以上即为 VTO 算法的全部步骤。利用 VTO 算法进行少数类样本生成的可视化效果如图 4 所示。

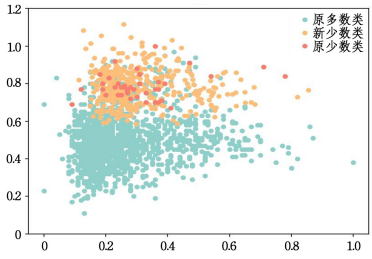


图 4 VTO 算法样本生成示意图

Fig. 4 VTO algorithm sample generation

2.3 算法时间复杂度

步骤 1 中，需要分别计算多数类与少数类的类中心点。已知多数类样本数目为 n_{major}，少数类样本数目为 n_{minor}，特征数为 n_f。由于涉及到对 $n_{\text{major}} + n_{\text{minor}}$ 个向量上的 n_f 个特征维度进行加法和除法操作，每个维度的加法和除法操作的时间复杂度都是 O(1)，因此整体的时间复杂度为 $O(n_{\text{major}} n_f + n_{\text{minor}} n_f)$。

步骤 2 中，需要计算用于迁移的多数类样本偏移向量。初始偏移向量的计算需要每个多数类样本减去多数类中心，时间复杂度为 $O(n_{\text{major}})$；计算多数类与少数类协方差矩阵的时间复杂度为 $O(n_{\text{major}} n_f^2)$；对协方差矩阵进行 PCA 提取因子载荷矩阵 Q 的时间复杂度小于 $O(n_f^3)$，计算新偏移向量的时间复杂度为 $O(n_f^3)$。

步骤 3 中，需要过滤掉不满足置信要求的迁移向量。由于多数类特征大于少数类的情况的计算更复杂，因此在研究时间成本时可仅考虑所有特征上多数类方差均大于少数类的情况。因此，首先对所有新偏移向量 $\widetilde{\Delta}_i$ 的每一个特征计算置信区间，该步的时间复杂度为 $O(n_{\text{major}} n_{\text{minor}} n_f)$；其次判断是否存在任一少数类样本，使得其每一个特征值都落在 $\widetilde{\Delta}_i$ 相应特征的置信区间内，该步的时间复杂度为 $O(n_{\text{major}} n_{\text{minor}}^2 n_f)$。

步骤 4 中，执行迁移，生成新样本。在少数类中心上叠加 N 个偏移向量，时间复杂度为 $O(N n_f)$。

从上述 4 个步骤的时间复杂度分析可知，VTO 算法的时间复杂度应由时间复杂度最大的部分表示，因此最终的时间复杂度为 $O(n_{\text{major}} n_{\text{minor}}^2 n_f)$。

3 实验结果与分析

3.1 数据集

为测试所提 VTO 算法的有效性，从 KEEL 公开数据库[15]中的非平衡板块选取 6 个数据集，详细信息如表 2 所列。从表 2 可以看出，在本文所选的数据集中，特征数最多为 18，最少为 7；非平衡度(Imbalance Ratio，IR)最高为 9.98，最低为1.86。这些数据集之间的差距较大，分类难度不同，可以有效验证本文算法的有效性和泛化性。

表 2 非平衡数据集详细信息

Table 2 Information of imbalanced data sets

数据集	特征数	样本数	IR
wisconsin	9	683	1.86
glass0	9	214	2.06
vehicle0	18	846	3.25
ecoli2	7	336	5.46
page-blocks0	10	5472	8.79
vowel0	13	988	9.98

此外，为了测试 VTO 算法在不同非平衡度下的有效性，本文额外选取平衡 KEEL 数据集 magic 进行随机抽取，构造不同 IR 的非平衡数据集，以探究在不同 IR 下该算法的效果差异。如表 3 所列，在 magic 数据集中，有 class 为 h 的总计 6688 例，定为正例样本；class 为 g 的总计 12332 例，定为负例样本；在其他参数不变的情况下，固定抽取正例样本数为 1000，随机抽取负例样本数量从 1000 递增至 10000，即构造 IR 从 1.0 逐步加到 10.0，步长为 0.2 的 45 个非平衡数据集。

表 3 随机抽取不同 IR 数据集详细信息

Table 3 Information of different IR data sets

数据集	特征数	正例数	负例数	IR
magic	10	1000	1000	1.0
			1200	1.2
			1400	1.4
			…	…
			9800	9.8
			10000	10.0

3.2 评价指标

F-score 是准确率和召回率的调和均值。准确率代表正确分类的样本占所有样本的比例，在不平衡分类中会受到多数类和少数类样本数量的影响，并不完全客观；而召回率代表少数类样本被正确分类的比例；只有当两者均增大时，才会使得 F-score 值提升，因此 F-score 值更能体现非平衡数据的分类效果。

PR 曲线刻画的是查准率和查全率之间的关系。查准率指的是在所有预测为正例的数据中真正例所占的比例，查全率是指预测为真正例的数据占所有正例数据的比例；当正负样本的比例失衡时，PR 曲线会产生很大的变化，因此适合用于评价非平衡数据集的分类效果。该曲线与坐标轴围成图形的面积值为 PR-AUC，因此本文以 F-score 和 PR-AUC 值为评价指标。

3.3 实验结果与分析

本文算法的编写与编译在 PyCharm 2022.1.3(Community Edition)中实现，实验环境为 AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz，16.0 GB RAM，64 位操作系统，Windows 11。

为了验证本文所提 VTO 算法的有效性，并与其他过采样方法进行对比，本文在不同的非平衡公开数据集上进行实验。实验首先对每个数据集进行 20 次 5 折交叉验证，总计 100 次训练结果，最后取所有结果的平均值。进行交叉验证时，取 20%的真实样本作为验证集，对剩余样本采用多种方法进行过采样后作为训练集；除本文将所提 VTO 外，对比算法包括经典算法 SMOTE 和 ASYDMN，以及改进的 Borderline-SMOTE，SMOTETomek，SMOTEENN 算法；过采样后使用决策树训练分类器并计算分类指标结果。

表4、表5分别展现了6个数据集上的F-score和PR-AUC实验结果。如表4所列,与未经过采样操作的origin数据集相比,VTO方法的F-score在6个数据集上分别提升了2.87%,3.99%,1.58%,1.22%,1.9%和4.01%,证明了本文所提方差迁移过采样方法的有效性;与其他过采样方法相比,VTO也具有相对较好的过采样性能,在4个数据集上取得最佳的分类效果,平均排名1.33,显著高于其余5种过采样方法。

表4 各过采样算法在不同数据集上的F-score值对比

Table 4 Comparison of F-score values of each oversampling algorithm on different data sets

数据集	origin	VTO	SMOTE	ADASYN	BD-SMOTE	SMOTETomek	SMOTEENN
wisconsin	0.9073	0.9360(2)	0.9248(4)	0.9172(5)	0.9133(6)	0.9250(3)	**0.9417(1)**
glass0	0.6853	**0.7252(1)**	0.7030(3)	0.6671(6)	0.6800(4)	0.6768(5)	0.7184(2)
vehicle0	0.8561	0.8719(2)	0.8671(4)	**0.8772(1)**	0.8680(3)	0.8630(5)	0.8504(6)
ecoli2	0.7604	**0.7726(1)**	0.7179(5)	0.7035(6)	0.7642(2)	0.7214(4)	0.7640(3)
page-blocks0	0.8182	**0.8372(1)**	0.8242(3)	0.8149(6)	0.8177(4)	0.8323(2)	0.8160(5)
vowel0	0.8927	**0.9328(1)**	0.8977(3)	0.9011(2)	0.8691(6)	0.8921(4)	0.8698(5)
平均排名	—	1.33	3.66	4.33	4.16	3.83	3.66

PR-AUC值的实验结果如表5所列。从表中可以看出,在6个数据集上VTO重采后的数据平均分类效果同样显著优于其他方法,进一步证明了本文提出的方差迁移方法能够提升过采样后的模型效果。

表5 各过采样算法在不同数据集上的PR-AUC值对比

Table 5 Comparison of PR-AUC values of each oversampling algorithm on different data sets

数据集	origin	VTO	SMOTE	ADASYN	BD-SMOTE	SMOTETomek	SMOTEENN
wisconsin	0.9297	0.9486(2)	0.9418(3)	0.9343(5)	0.9293(6)	0.9414(4)	**0.9499(1)**
glass0	0.7468	**0.7783(1)**	0.7574(3)	0.7258(6)	0.7308(5)	0.7371(4)	0.7631(2)
vehicle0	0.8720	0.8879(2)	0.8837(4)	**0.8895(1)**	0.8839(3)	0.8792(5)	0.8662(6)
ecoli2	0.7865	**0.8079(1)**	0.7443(5)	0.7337(6)	0.7997(2)	0.7605(4)	0.7867(3)
page-blocks0	0.8265	**0.8465(1)**	0.8338(3)	0.8235(6)	0.8263(5)	0.8409(2)	0.8324(4)
vowel0	0.9015	**0.9388(1)**	0.9040(3)	0.9076(2)	0.8749(6)	0.8990(4)	0.8770(5)
平均排名	—	1.33	3.5	4.33	4.5	3.83	3.5

此外,在VTO算法中,过采样的比例、方差迁移的半径阈值可能会对过采样效果有一定的影响。因此为了探究不同IR和不同参数与算法效果的关系,本文同样以决策树为基准分类器,随机以不同IR抽取公开数据集,在不同IR、不同参数条件下进行实验。

为探究VTO算法在不同IR数据集上的有效性,及IR对过采样效果的影响,本文在随机抽取出的45个IR递增的数据集上进行实验。图5、图6分别展现了在这些数据集上的F-score和PR-AUC实验结果,可见随着IR不断增加,VTO算法相对于其他过采样方法优势越加明显。

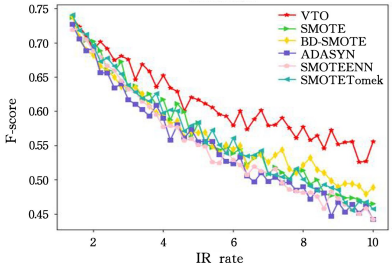


图5 不同IR下各过采样算法处理后的F-score值

Fig. 5 F-score of different oversample methods with different IR

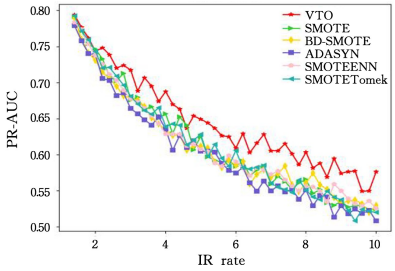


图6 不同IR下各过采样算法处理后的PR-AUC值

Fig. 6 PR-AUC of different oversample methods with different IR

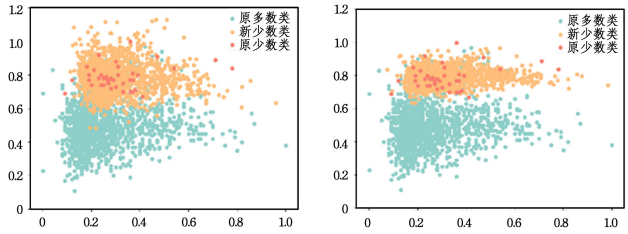
综上所述,本文提出的VTO算法分类性能良好,无论数据集的非平衡度如何递增,其都能相对原始数据和其他过采样方法有一定的提升。

3.4 不同参数对VTO算法效果的影响

VTO算法的效果受到以下3个参数的影响:主成分个数φ,参数K,以及置信概率阈值ρ。在实际代码实现时,一般以z来代替ρ以指定置信区间。

主成分个数φ用于控制协方差矩阵降维的程度。该值越大,新样本生成越分散;该值越小,因子载荷矩阵Q包含的信息越精炼,样本生成越集中,但也会有丢失特征信息的风险。

不同参数φ不同时的VTO算法的样本生成图如图7所示。



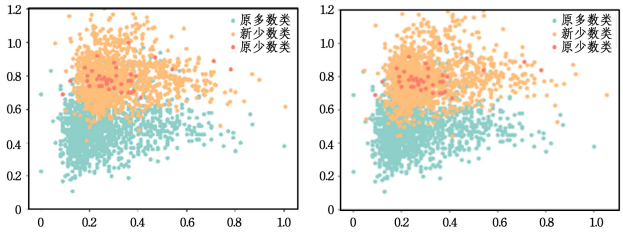
(a)$\varphi=3$ (b)$\varphi=2$

图7 参数φ不同时的VTO算法的样本生成图

Fig. 7 Samples generated by VTO algorithm with different φ

参数K决定了新样本生成时更多地服从多数类特征重要性排序抑或少数类特征重要性排序。其值越接近1,VTO算法越能从多数类中迁移信息;其值越接近0,新少数类样本生成越贴合少数类真实分布。

不同参数K不同时的VTO算法的样本生成图如图8所示。置信参数z控制新样本生成时的置信区域。z越小,所生成的新样本越靠近真实少数类样本点或少数类中心点。

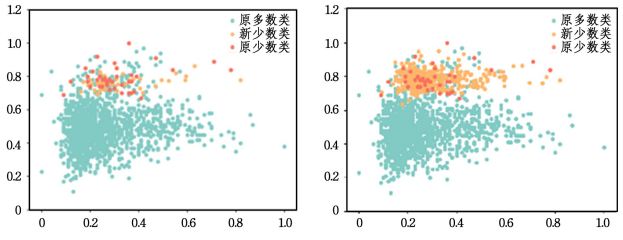


(a)$K=1$ (b)$K=0$

图 8 参数 K 不同时的 VTO 算样的本生成图

Fig. 8 Samples generated by VTO algorithm with different K

不同参数 z 下的 VTO 算法的样本生成图如图 9 所示。



(a)$z=0.65$ (b)$z=0.7$

图 9 参数 z 不同时的 VTO 算法的样本生成图

Fig. 9 Samples generated by VTO algorithm with different z

以上参数的最佳取值范围与数据集的具体情况息息相关，需多次重复实验以确定在各个数据集上效果最佳的参数组合。

结束语 为解决过采样过程中少数类信息不足的问题，提出了一种基于多数类方差迁移的少数类合成方法 VTO，从具有足够多样化样本的多数类中提取样本偏移量，综合少数类和多数类的特征权重矩阵，从而在少数类样本生成中引入类外信息，进而丰富少数类特征空间，促使少数类分布更接近常规分布。该方法突破了少数类由于样本过少导致的内含信息稀少的限制，使得过采样生成的新数据丰富了少数类类内信息，进而不易使分类模型发生过拟合，从而提升模型效果。实验决策树算法为分类模型在 6 个 KEEL 数据集上训练，对比 SMOTE 和 SMOTEENN 等其他过采样方法，以 F-score 和 PR-AUC 值作为评价指标。结果表明，VTO 可以对非平衡数据集进行有效的重平衡，在评价指标 F-score 和 PR-AUC 值上有明显的优势，是一种适用于不同非平衡度数据集的有效过采样方法。但其中参数的选择仍然需要通过多次实验进行确定，因此如何根据数据集自适应地适配参数，使生成样本更加符合真实分布，是未来的研究方向。

参 考 文 献

[1] ZHENG Y, WANG M. Imbalanced problem in initial coin offering fraud detection[C]// Proceedings of the Data Science. Singapore, 2022.

[2] CHEN L, XU G, ZHANG Q, et al. Learning deep representation of imbalanced SCADA data for fault detection of wind turbines [J]. Measurement, 2019, 139.

[3] ZENG M, ZOU B, WEI F, et al. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data [C]// IEEE International Conference of Online Analysis. 2016: 225-228.

[4] LIN X, CHEN Z, WANG Z. Aspect-level sentiment classification based on imbalanced data and ensemble learning [J]. Computer Science, 2022, 49(S1): 144-149.

[5] GUZMÁN-PONCE A, SÁNCHEZ J S, VALDOVINOS R M, et al. DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem [J]. Expert Systems with Applications, 2021, 168: 114301.

[6] JIN X, WANG L, SUN G, et al. Under-sampling Method for Unbalanced Data Based on Centroid Space [J]. Computer Science, 2019, 46(2): 50-55.

[7] KHUSHI M, SHAUKAT K, ALAM T M, et al. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data [J]. IEEE Access, 2021, 9: 109960-109975.

[8] CHAWLA N, BOWYER K, HALL L O, et al. SMOTE: Synthetic Minority Over-sampling Technique [J]. arXiv: 1106.1813, 2011.

[9] BARUA S, ISLAM M M, YAO X, et al. MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405-425.

[10] ASNIAR, MAULIDEVI N U, SURENDRO K. SMOTE-LOF for noise identification in imbalanced data classification [J]. Journal of King Saud University—Computer and Information Science, 2022, 34(6): 3413-3423.

[11] HAIRANI H, SAPUTRO K E, FADLI S. K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4. 5, SVM, dan naive Bayes [J]. Jurnal Teknologi dan Sistem Komputer. 2020: 5.

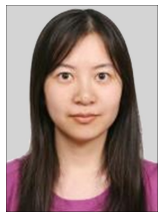
[12] ZHOU X, CAO F, YU L. Bi-directional oversampling method based on sample stratification [J]. Computer Science, 2019, 46(12): 83-88.

[13] ZHAO K, JIN X, WANG Y. Survey on few-shot learning [J]. Journal of Software, 2021, 32(2): 349-69.

[14] LIU J, SUN Y, HAN C, et al. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[15] ALCALÁ-FDEZ J, FERNÁNDEZ A, LUENGO J, et al. KEEL Data-Mining Software Tool: Data Set Repository[J]. Integration of Algorithms and Experimental Analysis Framework, 2011, 17: 255-287.

ZHENG Yifan, born in 2000, postgraduate. Her main research interests include fraud detection and imbalance data processing.



WANG Maoning, born in 1987, Ph. D, professor, is a member of the CCF(No. 93508M). Her main research interests include cryptography, blockchain and digital currency.