一种基于自适应子空间线性插值的数据合成方法

杜玉坤，金骁，汪红霞
（南京审计大学统计与数据科学学院，江苏 南京 211815）

摘　要：大数据时代，数据作为第五生产要素的基础，高质量的数据是军事智能化发展的重要保障，但是现实中的数据往往是含有大量未知噪音的，并且很难保重充足的样本以供智能技术的研究。本文提出了一种基于自适应子空间线性插值的数据合成方法，该方法可以在已有数据的基础上自适应地合成有效样本，并且可以有效降低数据平均噪音，调整样本结构，使得误差较大的样本占总样本比例显著下降，结合机器学习模型，可以显著提高模型的泛化能力。本文在理论上证明了该方法的有效性，模拟实验和实例分析的对比结果表示了该方法是有效且具有鲁棒性的。

关键词:自适应；未知噪音；数据合成；鲁棒性。
中图分类号：O242.1　　**文献标志码：**A

A data synthesis method based on adaptive subspace linear interpolation

Yukun Du, Xiao Jin, Hongxia Wang
(School of Statistics and Data Science, Nanjing Audit University, Nanjing, 211815, China)

Abstract: In the era of big data, where data serves as the foundation of the fifth factor of production, high-quality data is an essential guarantee for the development of military intelligence. However, real-world data often contains a large amount of unknown noise, and it's challenging to ensure sufficient samples for the study of intelligent technologies. This paper proposes a data synthesis method based on adaptive subspace linear interpolation. This method can adaptively synthesize valid samples based on existing data, effectively reduce the average noise of data, and adjust the sample structure, significantly reducing the proportion of samples with large errors. When combined with machine learning models, it can significantly improve the model's generalizability. This paper theoretically proves the effectiveness of this method, and the comparison results of simulation experiments and case analyses indicate that the method is both effective and robust.

Keywords: Adaptive; Unknown noise; Data synthesis; Robustness

0 引 言

在日常使用机器学习方法进行预测时，用于训练模型的数据经常会面临样本量不足、数据集观测误差较大等一系列问题。比如针对森林火灾燃烧面积预测，这种极小概率的突发事件可以收集到的样本信息较少，并且数据收集可能存在人为因素或不可控因素，例如设备误差、主观偏差、记录错误等，这些因素可能导致数据的质量下降。这就导致很难训练较高复杂度的模型以达到良好的预测效果。

目前比较流行的数据合成方法可以分为三类：第一类指在给定数据集上对已有的样本点之间进行插值或外推，比较具有代表性的有二次样条插值、三次样条差值[1]、多项式插值以及线性外推等方法[2-4]。对于图像数据，还可以使用变分自编码器（Variational Auto Encoder, VAE）和生成对抗网络（Generative Adversarial Networks, GAN）等一系列深度学习的相关方法来生成新的数据[5,6]。第二类方法主要是解决数据集不平衡的问题，例如 SMOTE[7]以及它的一些改进方法[8-10]。这些方法主要应用在分类问题中，可以合成少数类样本以平衡数据集并提高模型的性能。第三类方法主要针对敏感数据进行隐私保护，如差分隐私法[11]。

受到分段线性插值的启发，本文提出了一种基于插值思想的数据合成方法，该方法的思想是将原本的特征空间划分为具有等量样本的若干个子空间，然后对相邻子空间的样本通过特定的匹配方法进行线性插值。该方法可以自适应的扩充有效样本，并且可以调整噪音较大的样本占比，降低了数据噪声对预测结果的影响，达到优化样本的目的。本文将这种方法命名为自适应子空间线性插值(Adaptive Subspace Linear Interpolation, ASLI)。对于某些实际任务来说，原始数据通常包含大量特征和未知噪声[12]。ASLI 数据合成方法可以从两方面对样本进行优化：1、对含有噪音的原始数据集自适应地扩充了样本量，并且扩充生成的样本具有更小的噪声，降低了样本的平均噪声；2、自适应的调整样本结构，可以使得误差较大的样本占比减小。通过 ASLI 方法处理后的数据集,可以显著提高模型的预测效果，提升模型的泛化能力。

本文首先介绍了方法的整体思路以及具体的优化步骤，并提供了有效证明；第 2 节为模拟实验，对模拟数据进行优化并计算该方法的优化效果：第 3 节为实例应用，在实际数据集上使用 ASLI 方法结合机器学习模型进行预测，对比模型的泛化效果。

1 模型方法

ASLI 包含两个基础算法,即用于聚类的 K-Space 算法以及用于样本插值匹配的 K-Match 算法。本文首先对 ASLI 的整体流程进行介绍，然后再分别介绍 K-Space 与 K-Match 算法。

1.1 ASLI 方法整体介绍

ASLI 思想是将数据集原始特征空间划分为若干子空间，每个子空间包含相同数量的样本，然后对相邻特征子空间中的样本进行线性插值。该方法需要预先确定两个超参数（k 和 η），参数 k 的直观解释是每个特征子空间中存在的样本数，参数 η 指线性插值时单位距离插入的样本数量。

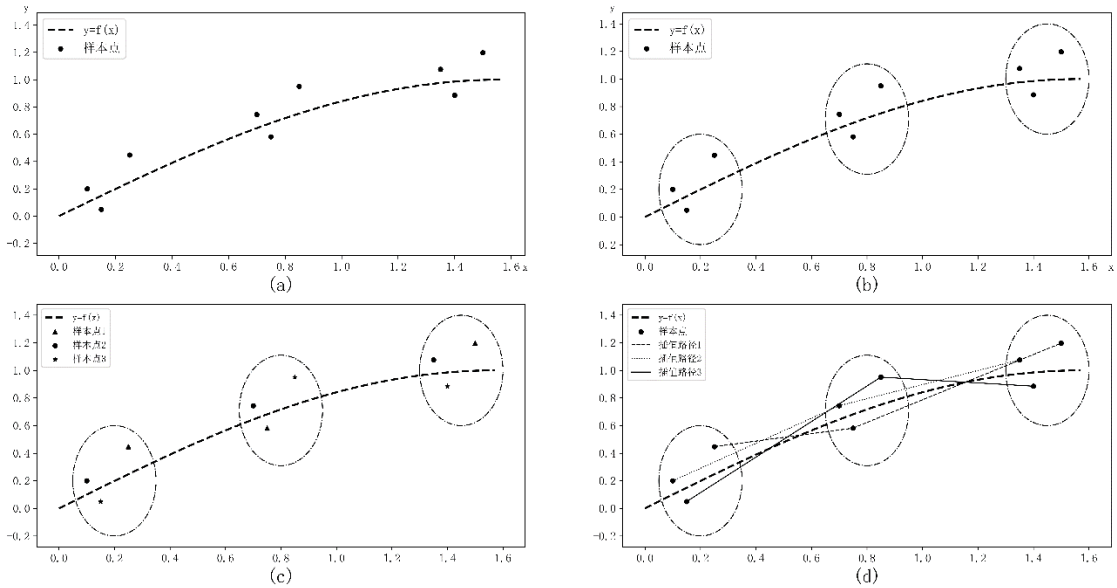

图 1 ASLI 方法过程

Fig. 1 ASLI method procedure

该方法流程图如图 1 所示。图 1(a)中展示了多个含有噪声的样本点，其中 $f(x)$ 为变量之间的真实函数关系；首先使用 K-Space 算法对样本点进行聚类，如图 1(b)所示；在图 1(c)中，使用 K-Match 算法对相邻子空间的样本进行样本插值匹配；最后对于匹配的样本之间进行线性插值，如图 1(d)所示。

对于给定数据集 $D=\{\boldsymbol{x}_i, y_i\}_{(i=1)}^{n}$，其中 $\boldsymbol{x}_i \in \mathrm{X}=\mathbf{R}^p$，$y_i \in \mathrm{Y} \equiv \mathbf{R}$。本文假设数据是含有噪声的，设 $\boldsymbol{x}_i$ 与的真实值为 $\dot{\boldsymbol{x}}_i$，$\dot{y}_i$。考虑 $\dot{y}_i=f(\dot{\boldsymbol{x}}_i)$,假设 $f(\cdot)$ 是一个连续的函数，表示 $\boldsymbol{x}_i$ 与 y_i 之间的真实函数关系。对于含噪数据，可得：

$$\dot{y}_i+\epsilon_{i,y}=f\left(\dot{\boldsymbol{x}}_i+\epsilon_{i,x}\right)+\epsilon_i, \tag{1}$$

其中，$\epsilon_{i,x}$,和 $\epsilon_{i,y}$ 分别为 $\boldsymbol{x}_i$ 与 y_i 的噪音，ϵ_i 为误差项。可将公式(1)转换为一种普遍的表示：

$$y_i=f(\boldsymbol{x}_i)+\epsilon_i. \tag{2}$$

令 $\boldsymbol{x}_0=\left\{x_0^1,\ldots,x_0^p\right\}$，对于 $\forall j=1,\ldots,p, x_0^j=\inf\{x_i^j\}_{i=1}^{n}$，称 $\boldsymbol{x}_0$ 为特征最小点。

给定超参数 k，使用无监督聚类 K-Space 算法对原始数据进行聚类。通过 K-Space 算法，可以将原始特征空间分为 n/k 个特征子空间(注:对于 n/k 不为整数的情况，后文会进行探讨

分析)，每个子空间包含等量的 k 个样本，即 X=$\bigcup_{s=1}^{n/k}$ X_s , $\mathrm{X}_i \cap \mathrm{X}_j = \varnothing, i, j = 1,2,\dots,\frac{n}{k}, i \neq j$，原始数据集可以根据每个子空间包含的样本划分为多个子集，即$D = \bigcup_{s=1}^{n/k} D_s, D_s = \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^k$， $\boldsymbol{x}_i^s \in \mathrm{X}_s$，如图 1(b)所示。对于两个相邻的子空间，由于$f(\cdot)$是连续的，因此可以假设在两个相邻子空间中，$f(\cdot)$可以拟合为一个线性函数$g(\cdot)$。公式(2)可以转化为:

$$y_i = g(\boldsymbol{x}_i) + \epsilon_i + \epsilon_i', \tag{3}$$

其中，ϵ_i'表示线性拟合误差.当子空间的测度以及相邻子空间的距离趋向于 0 时，可得$\epsilon_i' \to 0$。经过 K-Space 聚类之后，下一步需要对不同子空间对应的数据子集进行排序。

首先，需要计算各个簇的簇心，如下:

$$\overline{\boldsymbol{x}}^s = \frac{1}{k} \sum_{\boldsymbol{x}_i^s \in D_s} \boldsymbol{x}_i^s. \tag{4}$$

在集合$\{D_s\}_{s=1}^{n/k}$中，本文定义$D_{(1)}$为簇心距离特征最小点$\boldsymbol{x}_0$ 最近的子集，定义$D_{(d)}$为簇心距离$\overline{\boldsymbol{x}}^{(d-1)}$最近的子集，$\overline{\boldsymbol{x}}^{(d-1)}$为$D_{(d-1)}$的簇心，$D_{(d)} \neq D_{(1)}, \dots, D_{(d-1)}$，且$d > 1$。

$$D_{(1)} = \underset{D_s \in D}{\operatorname{argmin}}\, dist(\boldsymbol{x}_0, \overline{\boldsymbol{x}}^s), \tag{5}$$

$$D_{(d)} = \underset{\{D_s \in D,\, D_s \neq D_{(1)}, \dots, D_{(d-1)}\}}{\operatorname{argmin}} dist\left(\overline{\boldsymbol{x}}^{(d-1)}, \overline{\boldsymbol{x}}^s\right). \tag{6}$$

对排序后的子集$\{D_{(d)}\}_{d=1}^{n/k}$之间进行插值，序号相邻的子集为相邻的子空间对应的数据集，仅在相邻的子空间之间进行插值。

当对相邻子空间之间进行插值时，插值规则如下:

1、进行插值的两个样本属于不同的相邻子空间;

2、每个样本都要进行插值;

3、每个样本只能进行一次插值。

根据以上规则，可选择的插值方案有 k!个，并不是所有的方案插值后都能达到很好的优化效果，因此本文通过 K-Match 样本匹配算法对相邻子空间当中的 2k 个样本进行线性插值匹配，该算法可以从 k!个方案中快速且高效的选择优化效果较好的插值方案$\{(\boldsymbol{x}_i^{(d)}, y_i^{(d)}), (\boldsymbol{x}_i^{(d+1)}, y_i^{(d+1)})\}_{i=1}^k$。

假设$\boldsymbol{x}$和y都是连续的变量，给定另外一个超参数 η，对相邻子空间$D_{(d)}$与$D_{(d+1)}$包含的样本进行线性插值，插入的样本数量为$\sum_{i=1}^{\mathrm{k}} \left\lfloor \eta \cdot dist\left(\boldsymbol{x}_i^{(d)}, \boldsymbol{x}_i^{(d+1)}\right) \right\rfloor$。以$\left(\boldsymbol{x}_i^{(d)}, y_i^{(d)}\right) \in D_{(d)}$ 和$\left(\boldsymbol{x}_i^{(d+1)}, y_i^{(d+1)}\right) \in D_{(d+1)}$为例，插值生成的样本数据集为$\{(\boldsymbol{x}_{(d,d+1)}^{(m,i)}, y_{(d,d+1)}^{(m,i)})\}_{m=1}^{\left\lfloor \eta \cdot dist\left(\boldsymbol{x}_i^{(d)}, \boldsymbol{x}_i^{(d+1)}\right) \right\rfloor}$。插值公式为:

$$\boldsymbol{x}_{(d,d+1)}^{(m,i)} = \boldsymbol{x}_i^{(d)} + m \cdot \frac{\boldsymbol{x}_i^{(d+1)} - \boldsymbol{x}_i^{(d)}}{\left\lfloor \eta \cdot dist\left(\boldsymbol{x}_i^{(d)}, \boldsymbol{x}_i^{(d+1)}\right)\right\rfloor + 1}, \tag{7}$$

$$y_{(d,d+1)}^{(m,i)} = y_i^{(d)} + m \cdot \frac{y_i^{(d+1)} - y_i^{(d)}}{\left\lfloor \eta \cdot dist\left(y_i^{(d)}, y_i^{(d+1)}\right)\right\rfloor + 1}. \tag{8}$$

ASLI 方法的主要处理过程见算法 1：

算法 1 ASLI 算法

输入: 原始数据集 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, 超参数 k 和 η

输出: 优化后的数据集 $D^{'}$

1: 通过 K-Space 算法对原始数据集进行聚类处理，得 $D = \{D_s\}_{s=1}^{n/k}$，其中$D_s = \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{k}$

2: 计算每个簇的簇心：$\{\overline{\boldsymbol{x}}^s\}_{s=1}^{n/k}$

3: 根据公式(5)和(6)对多个簇进行排序，可得：$\{D_{(d)}\}_{d=1}^{n/k}$

4: 令 $d = 1, \ldots, \frac{n}{k} - 1$ 进行迭代循环

5: 通过 K-Match 算法对数据子集 $D_{(d)}$ 和 $D_{(d+1)}$ 的样本进行匹配，得: $\{(\boldsymbol{x}_i^{(d)}, y_i^{(d)}), (\boldsymbol{x}_i^{(d+1)}, y_i^{(d+1)})\}_{i=1}^{k}$

6: 令 $i = 1, \ldots, k$ 进行迭代循环

7: 通过公式(7)和(8)计算合成新样本 $\{(\boldsymbol{x}_{(d,d+1)}^{(m,i)}, y_{(d,d+1)}^{(m,i)})\}_{m=1}^{\lfloor \eta \cdot dist(x_i^{(d)}, x_i^{(d+1)})\rfloor}$

8: $D^{'} \leftarrow D \cup \{(\boldsymbol{x}_{(d,d+1)}^{(m,i)}, y_{(d,d+1)}^{(m,i)})\}_{m=1}^{\left\lfloor \eta \cdot dist\left(x_i^{(d)}, x_i^{(d+1)}\right)\right\rfloor}$

综上，对数据集进行 ASLI 处理,数据集应当满足的假设如下:

1、$f(\cdot)$是一个连续的函数;

2、线性拟合误差 $\epsilon_i \to 0$;

3、$\boldsymbol{x}$ 和 y 是连续的变量。

1.2 K-Space 聚类算法

ASLI 算法的实现首先需要将特征空间划分为具有等量样本的多个子空间，每个子空间含有 k 个样本，即$D = \bigcup_{s=1}^{n/k} D_s$, $D_s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{k}$，并且$D_i \cap D_j = \emptyset, i \neq j$。基于此目的,本文提出了 K-Space 无监督聚类算法。

对于特征子空间 X_s 所对应的数据集D_s，首先根据公式(9)计算得到$\boldsymbol{x}_1^s$：

$$\boldsymbol{x}_1^s = \underset{\boldsymbol{x}:\boldsymbol{x}\in D, \boldsymbol{x}\notin D_1, \ldots, D_{s-1}}{\operatorname{argmin}} dist(\boldsymbol{x}, \overline{\boldsymbol{x}}^{s-1}), \tag{9}$$

其中，$s=1,\ldots,\frac{n}{k}$, $\bar{\boldsymbol{x}}^{s-1}$是$D_{s-1}$的簇心且$\bar{\boldsymbol{x}}^{0}=\boldsymbol{x}_0$。定义$D_s=\{\boldsymbol{x}_1^s\}$，根据公式(10)计算可得$\boldsymbol{x}_d^s$：

$$\boldsymbol{x}_d^s=\underset{\boldsymbol{x}:\boldsymbol{x}\in D,\boldsymbol{x}\notin D_1,\ldots,D_s}{argmin}\ dist(\boldsymbol{x},\bar{\boldsymbol{x}}^s), \tag{10}$$

其中，$d=2,\ldots,k$。更新数据集$D_s\leftarrow D_s\cup\{\boldsymbol{x}_d^s\}$。

算法 2 K-Space 无监督聚类算法

输入: 原始数据集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$, 超参数$k$

输出: 聚类后具有等量样本的多个子集$\{D_s\}_{s=1}^{n/k}$

1: 计算特征最小点 $\boldsymbol{x}_0$, $\forall j=1,\ldots,p, x_0^j=\inf\{x_i^j\}_{i=1}^n$

2: 令 $s=1,\ldots,n/k$ 进行迭代循环

3: $\quad\boldsymbol{x}_1^s=\underset{\boldsymbol{x}:\boldsymbol{x}\in D,\boldsymbol{x}\notin D_1,\ldots,D_{s-1}}{\operatorname{argmin}}\ dist(\boldsymbol{x},\bar{\boldsymbol{x}}^{s-1})$

4: $\quad D_s=\{\boldsymbol{x}_1^s\}$

5: $\quad$令 $d=2,\ldots,k$ 进行迭代循环

6: $\quad\quad\boldsymbol{x}_d^s=\underset{\boldsymbol{x}:\boldsymbol{x}\in D,\boldsymbol{x}\notin D_1,\ldots,D_s}{argmin}\ dist(\boldsymbol{x},\bar{\boldsymbol{x}}^s)$

7: $\quad\quad D_s\leftarrow D_s\cup\{\boldsymbol{x}_d^s\}$

1.3 K-Match 样本匹配算法

令$\mathrm{X}=\mathbf{R}$，在对相邻子空间的样本进行线性插值时,可以计算得出每个插值匹配方案的总误差为:

$$\sum_{i=1}^{k}S(x_i^{(d)},x_i^{(d+1)})=\sum_{i=1}^{k}\int_{x_i^{(d)}}^{x_i^{(d+1)}}|f(x)-L_i(x)|\mathrm{d}x, \tag{11}$$

其中，$L_i(x)$为经过$(x_i^{(d)},y_i^{(d)})$和$(x_i^{(d+1)},y_i^{(d+1)})$两点的直线。

定理 1:对于两个相邻的子空间$\mathrm{X}_{(d)}$和$\mathrm{X}_{(d+1)}$对应的子集$D_{(d)},D_{(d+1)}$。$(\boldsymbol{x}_i^{(d)},y_i^{(d)})\in D_{(d)}$，$(\boldsymbol{x}_i^{(d+1)},y_i^{(d+1)})\in D_{(d+1)}$。函数关系为$y_i=f(x_i)+\epsilon_i$，则$\epsilon_i^{(d)}=y_i^{(d)}-f(x_i^{(d)})$。对于$\forall i=1,2,\ldots,k$，假设$\epsilon_i'\to 0$，则有$E(\frac{S(x_i^{(d)},x_i^{(d+1)})}{|x_i^{(d)}-x_i^{(d+1)}|})<E(\frac{(|\epsilon_i^{(d)}|+|\epsilon_i^{(d+1)}|)}{2})$。

证明 1：由于$\epsilon_i'\to 0$，根据公式(3)可得：

$$y_i=g(x_i)+\epsilon_i,$$

其中，$g(\cdot)$为线性函数。根据公式(11)可得：

$$S\left(x_i^{(d)},x_i^{(d+1)}\right)=\int_{x_i^{(d)}}^{x_i^{(d+1)}}|g(x)-L_i(x)|\mathrm{d}x,$$

当$\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0$时，令$(x',y')$为直线$y=L_i(x)$和$y=g(x)$的交点，我们可以简化$S(x_i^{(d)},x_i^{(d+1)})$的计算，并且根据迭代期望定律(LIE)可得:

$$E\left(\frac{S\left(x_i^{(d)},x_i^{(d+1)}\right)}{\left|x_i^{(d)}-x_i^{(d+1)}\right|}\right)=$$

$$\frac{E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}>0\right)P\left(\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}\geq 0\right)+E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)P\left(\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)}{|x_i^{(d)}-x_i^{(d+1)}|}=$$

$$\frac{E(|\epsilon_i^{(d)}|+|\epsilon_i^{(d+1)}|)P\left(\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}\geq 0\right)+\left(h_1\cdot E\left(|\epsilon_i^{(d)}|\right)+h_2\cdot E\left(|\epsilon_i^{(d+1)}|\right)\right)P\left(\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)}{2},$$

其中$h_1=\frac{|x'-x_i^{(d)}|}{|x_i^{(d)}-x_i^{(d+1)}|}$，$h_2=\frac{|x'-x_i^{(d+1)}|}{|x_i^{(d)}-x_i^{(d+1)}|}$。由于$P\left(\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}\geq 0\right)+P\left(\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)=1$，并且$h_1+h_2=1$，可得$E(\frac{S(x_i^{(d)},x_i^{(d+1)})}{|x_i^{(d)}-x_i^{(d+1)}|})<E(\frac{(|\epsilon_i^{(d)}|+|\epsilon_i^{(d+1)}|)}{2})$。

可以通过定理 1 得到，如果完全随机的从 k!个插值匹配方案中选择一个方案进行特征子空间插值处理，其均匀误差的期望是小于原始数据集均匀误差期望。但是随机选择并不能达到很好的优化效果，也不能保证结果的唯一性。从定理 1 的证明可以看出，对$x_i^{(d)}$和$x_i^{(d+1)}$之间进行插值，如果$\varepsilon_i^{(d)}\cdot\varepsilon_i^{(d+1)}<0$，则生成的样本具有更小的噪音。

定理 2: 令$y_i^{(d)}=f\left(x_i^{(d)}\right)+\epsilon_i^{(d)}$，$y_i^{(d+1)}=f\left(x_i^{(d+1)}\right)+\epsilon_i^{(d+1)}$。假设$\epsilon_i'\to 0$，可得：

$$E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)\middle|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)<E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)\middle|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}\geq 0\right)。$$

证明 2: 由于 $\epsilon_i'\to 0$, 基于定理 1 的证明，可得：

$$E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)\middle|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}\geq 0\right)=\frac{\left|x_i^{(d)}-x_i^{(d+1)}\right|E(|\epsilon_i^{(d)}|)+|x_i^{(d)}-x_i^{(d+1)}|E|\epsilon_i^{(d+1)}|}{2},$$

$$E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)\middle|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)=\frac{|x'-x_i^{(d)}|\cdot E\left(\left|\epsilon_i^{(d)}\right|\right)+|x'-x_i^{(d+1)}|\cdot E\left(\left|\epsilon_i^{(d+1)}\right|\right)}{2}.$$

综上可得$E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)\middle|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}<0\right)<E\left(S\left(x_i^{(d)},x_i^{(d+1)}\right)\middle|\epsilon_i^{(d)}\cdot\epsilon_i^{(d+1)}\geq 0\right)$。

根据定理 2 可得，噪音异号的样本之间进行插值生成的样本具有更小的噪音，因此 K-Match 算法的核心思想是判断每个样本的噪音正负，并且尽可能使噪音异号的样本之间进行插值。在 K-Match 算法中，首先应当根据数据集中噪音分布的情况选择合适的线性回归方法拟合数据集$D_{(d)}\cup D_{(d+1)}$得到线性函数$\hat{g}(\cdot)$。线性回归方法可以选择 Lasso 回归，局部加权线性回归(Locally Weighted Linear Regression，LWLR)[13]。或者具有鲁棒性的其他线性回归方法[14-15]。在后文的模拟实验以及实例分析中，本文主要采用的线性拟合方法为 OLS 以及支持向量回归(Support Vector Regression，SVR)。

根据公式(3)，假设线性拟合误差$\epsilon_i'\to 0$，对于数据集$D_{(d)}\cup D_{(d+1)}$，可得:

$$\epsilon_i=y_i-\hat{g}(\boldsymbol{x}_i). \tag{12}$$

根据ϵ_i的值，对$D_{(d)}$的样本进行正向排序可得$\left\{\left(\boldsymbol{x}_i^{(d)},y_i^{(d)}\right)\right\}_{i=1}^{k}$，对于$D_{(d+1)}$中的样本进行逆向排序可得$\left\{\left(\boldsymbol{x}_i^{(d+1)},y_i^{(d+1)}\right)\right\}_{i=1}^{k}$。对分别位于$D_{(d)}$以及$D_{(d+1)}$中相同序号的样本进行插值匹配，可得匹配方案$\{(\boldsymbol{x}_i^{(d)},y_i^{(d)}),(\boldsymbol{x}_i^{(d+1)},y_i^{(d+1)})\}_{i=1}^{k}$。

算法 3 K-Match 样本匹配算法

输入: 数据集$D_{(d)}$，$D_{(d+1)}$

输出: 样本插值匹配方案$\{(\boldsymbol{x}_i^{(d)}, y_i^{(d)}), (\boldsymbol{x}_i^{(d+1)}, y_i^{(d+1)})\}_{i=1}^{k}$

1: 对数据集$D_{(d)} \cup D_{(d+1)}$进行线性拟合,得$\hat{g}(\cdot)$

2: 根据公式(12)计算得$\{\epsilon_i\}$.

3: 对数据集$D_{(d)}$，$D_{(d+1)}$中得样本分别进行正向或逆向排序

4: 根据排序结果得到插值匹配方案$\{(\boldsymbol{x}_i^{(d)}, y_i^{(d)}), (\boldsymbol{x}_i^{(d+1)}, y_i^{(d+1)})\}_{i=1}^{k}$

1.4 方法补充

对于满足 ASLI 假设的数据集，该方法可以有效的扩充样本量，调整数据集的样本结构，降低样本与实际分布的均匀误差.该方法的直观处理效果如图 2 所示：

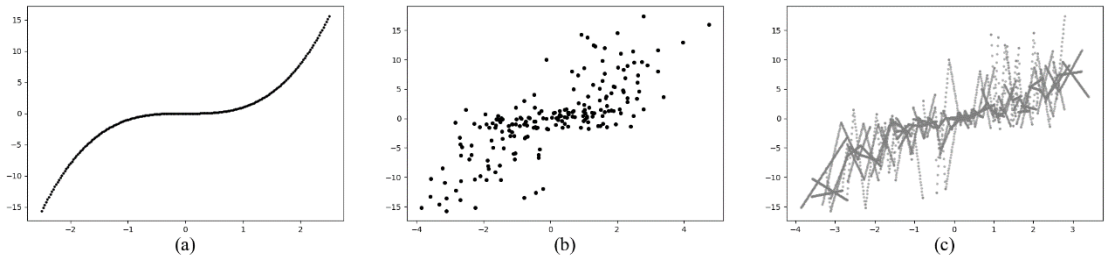

图 2 ASLI 处理效果图

Fig. 2 ASLI Processing Results

如图 2 所示。图 2(a)中，生成实际函数关系为$y = x^3$的样本，样本量为 200；对样本添加高斯噪声，见图 2(b)；图 2(c)中，令 k=6，η=100，对噪音数据使用 ASLI 方法进行数据合成处理，样本量增加至 3808。

在 ASLI 方法中，如果数据的特征之间的量纲存在显著差异，数量级较大的特征维度会导致生成的样本较为冗余，应当对数据进行归一化处理。许多情况下，n/k 并不是一个整数，针对这种情况可以使用 LOF 离群点检测算法[16]，通过该算法检测出多余数量的离群样本点，这些样本点将不参与 ASLI 处理。

2 模拟实验

2.1 评价指标及数据准备

本节主要探究在模拟数据集中，ASLI 方法的数据合成优化效果。由于 ASLI 方法可以降低数据的均匀噪音，并且可以降低较大误差样本所占比例，本文通过计算经过 ASLI 方法处理前后的样本与实际分布的均方误差(MSE)以及不同误差的样本占比作为评判指标，如公式(13)，(14)所示:

$$p(\alpha) = \frac{1}{n}\sum_{i=1}^{n} I(\left|f(\boldsymbol{x}_i) - y_i\right| > \alpha), \tag{13}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2. \tag{14}$$

生成模拟数据$\{\boldsymbol{x_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sim N_p(\mathbf{0}, \mathbf{E})$，生成权重矩阵$\boldsymbol{W_1} \in \mathbb{R}^{(p \times p_1)}$，$\boldsymbol{W_2} \in \mathbb{R}^{(p_1 \times 1)}$，$\boldsymbol{W_1}$,$\boldsymbol{W_2}$所有参数相互独立且服从高斯分布。$y_i = f(\boldsymbol{x_i}) + \epsilon_i = \tanh(\boldsymbol{x'_i W_1})\boldsymbol{W_2} + \epsilon_i$。

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{15}$$

对于给定的含有噪音的原始数据集，通常难以确定其噪音的真实分布[17]。因此，本文将服从均匀分布的噪音与多个正态分布的噪音按照一定权重混合，构建了包含未知噪音的数据集[18]。共生成 6 组实验数据，详见表 1。本文所有实验均固定了随机效应。

表 1 模拟数据生成

Table 1 Simulated data generation

模拟数据集	ϵ_i 分布	样本量	(P, P_1)
D_1	20%-N (0,64) 30%-U (-8,8) 50%-N (0,0.04)	500	(5,3)
D_2	20%-N (0,64) 30%-U (-8,8) 50%-N (0,0.04)	200	(5,3)
D_3	20%-N (0,64) 30%-U (-8,8) 50%-N (0,0.04)	1500	(5,3)
D_4	20%-N (0,64) 30%-U (-8,8) 50%-N (0,0.04)	500	(1,3)
D_5	20%-N (0,64) 30%-U (-8,8) 50%-N (0,0.04)	500	(20,10)
D_6	40%-N (0,64) 45%-U (-8,8) 15%-N (0,0.04)	500	(5,3)

2.2 超参数取值分析

本文首先探究超参数 k 的取值，令 $\eta = 1$, $k = 1,2,\ldots,\left\lfloor\frac{n}{2}\right\rfloor$。计算不同 k 的取值下，数据集经过 ASLI 方法处理前后的 MSE，并求得$k' = \underset{k:k=1,2,\ldots,\left\lfloor\frac{n}{2}\right\rfloor,\eta=1}{\operatorname{argmin}} \mathrm{MSE}$。超参数 k 不同取值下，

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 , \tag{16}$$

MSE 的变化如图 3 所示

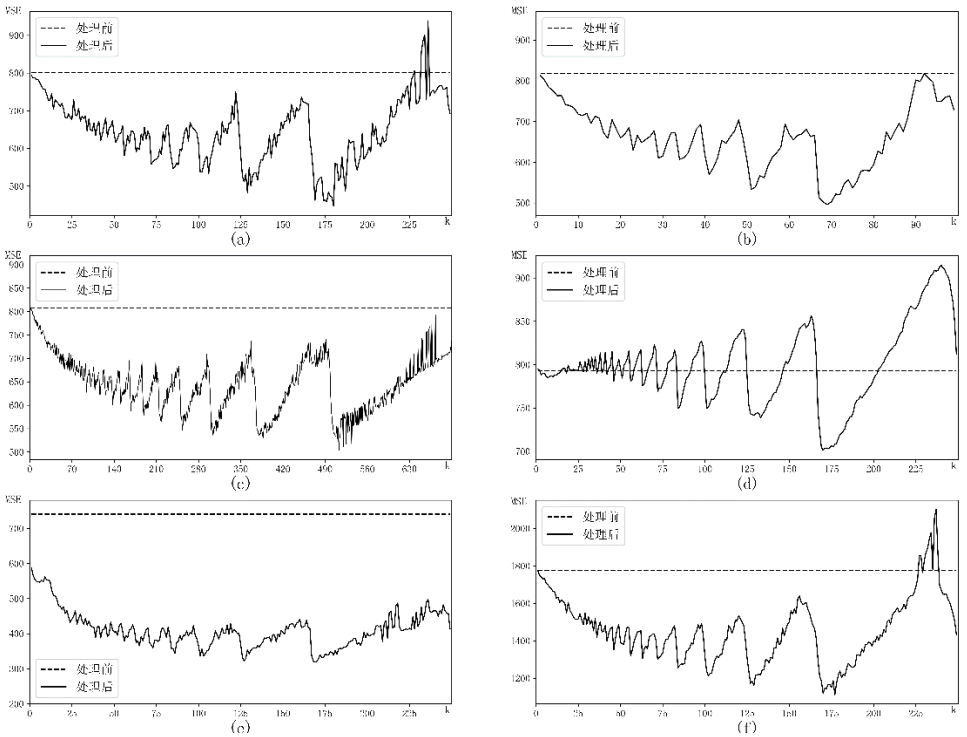

图 3 不同超参数 k 值下的 MSE

Fig.3 Changes in the MSE under different k values

令$\eta = 1,2,\ldots,30$，$k = k'$，计算数据集经过 ASLI 方法处理前后的 MSE，并求得$\eta' = \underset{\eta:\eta = 1,2,\ldots,30, k=k'}{\operatorname{argmin}} \text{MSE}$，如图 4 所示：

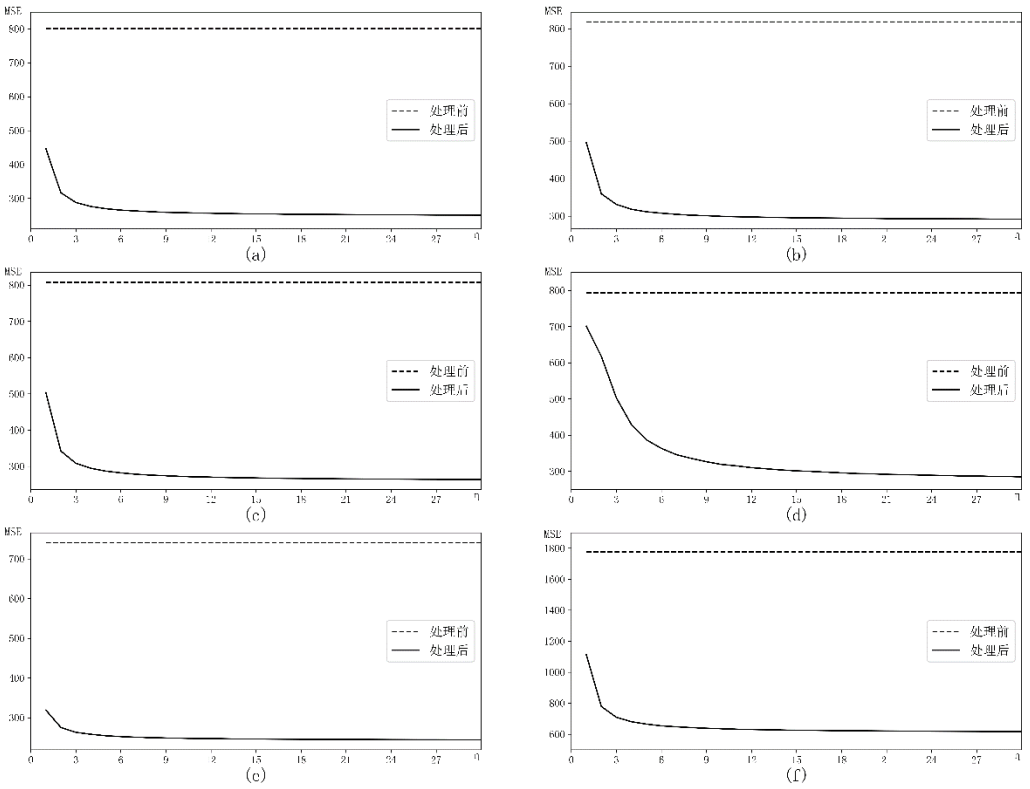

图 4 不同超参数η取值下的 MSE

Fig.4 Changes in the MSE under different η values

最后，令$k = k'$, $\eta = 1, \eta'$，计算不同α下，经过 ASLI 处理前后指标$p(\alpha)$的优化效果:

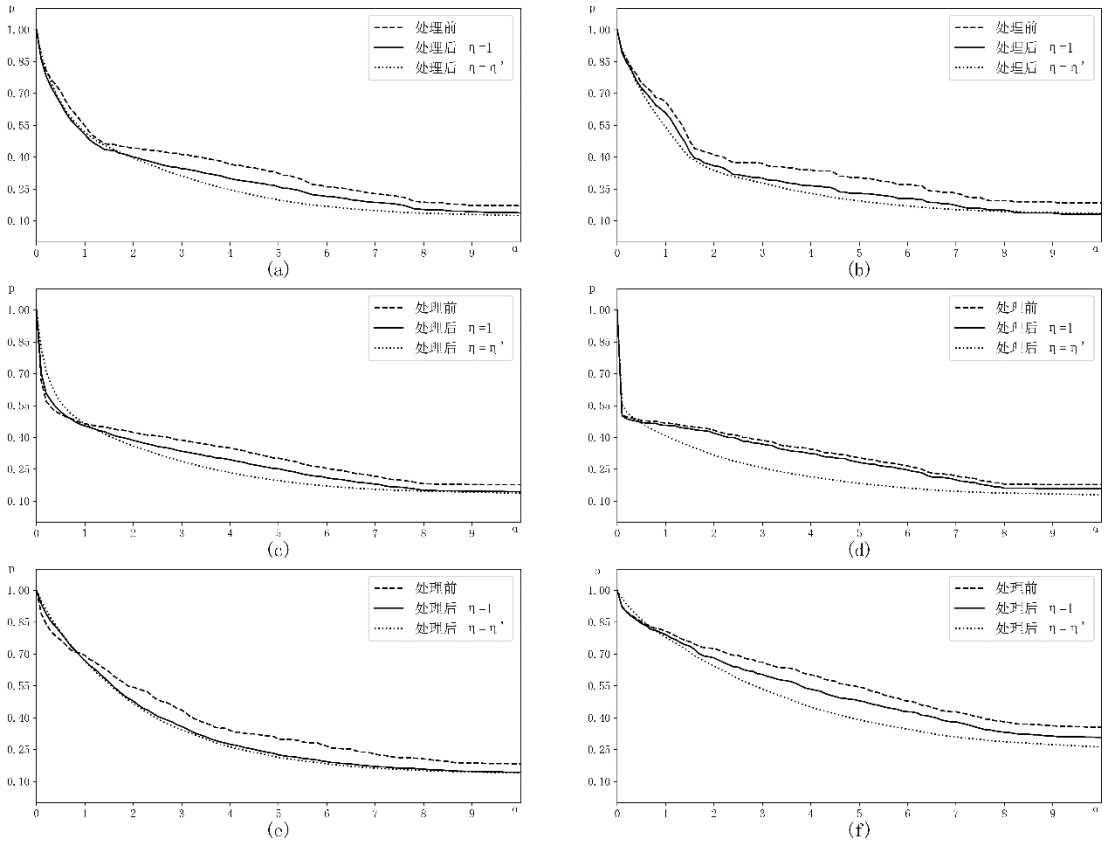

图 5 不同η取值下的 $p(\alpha)$

Fig.5 Changes in the p(α) under different η values

图 3 实验结果表明，随着大方差噪声含量的增加，ASLI 的性能不会显著下降,它能够更好地处理未知噪声，并具有良好的鲁棒性。从图 4 可以看出，超参数η与 MSE 一般呈单调递减关系，η的值越大，影响越显著。从图 5 中可以看出，ASLI 可以自适应地调整样本结构，从而降低了误差较大的样本比例。

2.3 方法对比

将 ASLI 方法与其他插值方法进行对比,对比方法选择:分段线性插值,线性外推法以及最近邻插值法. 基于给定的数据集，本文使用 ASLI 生成的样本作为插值点来计算线性外推和最近邻插值中的输出值。此外，取 ASLI 的超参数 k=1，$\eta = \eta'$，作为分段线性插值方法进行数据合成。实验结果表明，ASLI 的 MSE 最小。

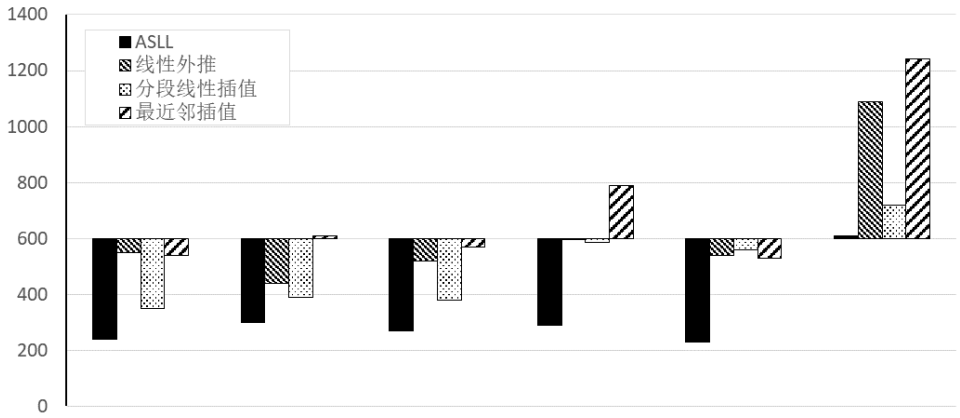

图 6 优化效果方法对比

Fig. 6 Comparison of optimization methods

3 实例应用

使用 Bike Sharing Demand、Air Quality、Forest Fires 三个实例数据[19-20]，按照 7:3 的比例划分训练集与测试集，并进行极大极小归一化处理。对训练集数据使用 ASLI 方法进行数据合成，机器学习预测模型选择 K 近邻(K-Nearest Neighbor, KNN)，随机森林(Random Forest, RF)，梯度提升决策树(Gradient Boosting Decision Tree, GBDT)，多层感知机(Multilayer Perceptron, MLP)以及支持向量回归机(Support Vector Regression, SVR)。评价指标选择均匀绝对误差(Mean Absolute Error, MAE)。预测效果详见表 2：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \qquad (17)$$

表 2 实例数据预测效果

Table2 Prediction Performance on Real Data

数据集	处理	超参数	训练集样本量	测试集 MAE				
				KNN	RF	MLP	SVR	GBDT
Bike Sharing Demand	-	-	7620	0.022	**0.0012**	0.0054	0.0426	0.0023
	ASISO	K=150, η=10	40380	0.020	**0.0006**	0.0025	0.0412	0.0019
Air Quality	-	-	6549	0.0299	**0.0257**	0.0480	0.0387	0.0258
	ASISO	K=20, η=100	101984	0.0289	0.0255	**0.0208**	0.0384	0.0277
Forest Fires	-	-	361	**0.0406**	0.0493	0.1058	0.0736	0.0449
	ASISO	K=10, η=100	21131	**0.0389**	0.0432	0.048	0.1005	0.0414

通过表 2 可知，每个数据集经过 ASLI 方法处理后,都可以有效提升模型的泛化能力。并且，以上数据集包含许多非连续变量，这并不满足 ASLI 方法的假设，以及对于样本量较少

的数据集(如 Forest Fires)，在子空间之间插值时，很难保证线性拟合误差$\epsilon_i^{'} \to 0$。这表明，即使在实际应用中存在违反 ASLI 假设的情况，该方法仍可能取得较好的优化结果。

4 结论

本文提出了一种数据合成方法 ASLI，它可以自适应地调整数据集的大小，且扩展后的数据通常包含最小的实际错误。此外，它还可以调整样本的结构，这可以显著降低大误差样本的比例。对模拟数据集的实验结果表明，ASLI 可以优化样本，相比于其他方法，使用这种方法生成的数据具有更小的误差，并且能更好地处理未知噪声，具有良好的鲁棒性。在实例数据集上的实验结果显示，该方法可以适用于多种机器学习模型，并且在大多数情况下，可以提高模型的泛化能力。

参考文献

[1] 陈娟,李崇君.基于 SBFEM 和样条插值的多边形偶应力/应变梯度理论单元[J].中国科学:物理学 力学 天文学,2021,51(05):135-150.
CHEN Juan, LI Chongjun. Polygon couple stress/strain gradient theory element based on SBFEM and spline interpolation [J]. Scientia Sinica Physica, Mechanica & Astronomica, 2021, 51(05): 135-150.

[2] 高强,高敬阳,赵地.GNNI U-net:基于组归一化与最近邻插值的 MRI 左心室轮廓精准分割网络[J].计算机科学,2020,47(08):213-220.
GAO Qiang, GAO Jingyang, ZHAO Di. GNNI U-net: An accurate segmentation network for MRI left ventricular contour based on group normalization and nearest neighbor interpolation [J]. Computer Science, 2020, 47(08): 213-220.

[3] 高兴华,李宏,刘洋.非线性分数阶常微分方程的分段线性插值多项式方法[J].应用数学和力学,2021,42(05):531-540.
GAO Xinghua, LI Hong, LIU Yang. A piecewise linear interpolation polynomial method for nonlinear fractional-order ordinary differential equations [J]. Applied Mathematics and Mechanics, 2021, 42(05): 531-540.

[4] 齐宗会,汪晖,刘永平.连续可导函数类的最优拉格朗日插值[J].高等学校计算数学学报,2020,42(01):87-96.
QI Zonghui, WANG Hui, LIU Yongping. Optimal Lagrange interpolation for continuously differentiable function classes [J]. Journal of Computational Mathematics in Colleges and Universities, 2020, 42(01): 87-96.

[5] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 2021, 65(5), 545–563.

[6] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data, 2019, 6(1): 1–48.

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002, 16, 321–357.

[8] Dablain, D., Krawczyk, B., Chawla, N.V.: Deepsmote: Fusing deep learning and smote for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 2022

[9] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International conference on intelligent computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 878-887.

[10] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic

minority over-sampling technique for handling the class imbalanced problem[C]//Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13. Springer Berlin Heidelberg, 2009, 475-482.

[11] Ha, T., Dang, T.K., Dang, T.T., Truong, T.A., Nguyen, M.T.: Differential privacy in deep learning: an overview. In: 2019 International Conference on Advanced Computing and Applications (ACOMP)IEEE, 2019, 97–102.

[12] Meng, D., De La Torre, F.: Robust matrix factorization with unknown noise.In: Proceedings of the IEEE International Conference on Computer Vision, 2013, 1337–1344.

[13] Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association, 1979, 74(368): 829–836.

[14] Lichti, D.D., Chan, T.O., Belton, D.: Linear regression with an observation distribution model. Journal of geodesy, 2021, 95: 1–14.

[15] Liu, C., Li, B., Vorobeychik, Y., Oprea, A.: Robust linear regression against training data poisoning. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, 91–102.

[16] Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, 93–104.

[17] Meng, D., De La Torre, F.: Robust matrix factorization with unknown noise.In: Proceedings of the IEEE International Conference on Computer Vision, 2013, 1337–1344.

[18] Guo, Y., Wang, W., Wang, X.: A robust linear regression feature selection method for data sets with unknown noise. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 31–44

[19] Cukierski, W.: Bike Sharing Demand, Kaggle, 2014. https://kaggle.com/competitions/bike-sharing-demand

[20] Dua, D., Graff, C.: UCI Machine Learning, Repository, 2017. http://archive.ics.uci.edu/ml

作者简介（必填项）

第一作者姓名：杜玉坤（1997-），性别男，学历硕士，主要研究方向为数据挖掘、机器学习、复杂系统。

第二作者姓名:金骁（1998-），性别男，学历硕士，主要研究方向为机器学习、多任务学习、非参数估计、时空数据分析。

第三作者姓名：汪红霞（1983-），性别女，职称副教授，学历博士，主要研究方向多任务学习、**非参数统计分析**、时空数据分析、数字经济。