An adaptive multi-path linear interpolation method for sample optimization

Abstract

在我们通过机器学习方法进行预测时，经常会面临样本量较少或者观测样本噪音较大的问题，现主流的样本扩充方法并不能很好的处理数据噪音问题。我们提出了一种基于线性插值思想的多路径样本扩充方法（AMLI），该方法主要解决预测样本量不足或观测样本与实际分布误差较大的问题。AMLI 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，将每个子空间各随机提取一个样本作为一个类，然后对于同类中的样本进行线性插值，既 K-path linear interpolation。经过 AMLI 方法处理后，可以极大程度扩充有效样本，调整样本结构，降低样本的平均噪音，以此来提升机器学习模型的预测效果。该方法的超参数具有直观的解释，通常需要很少的调优，本文结合了多种机器学习预测方法，结果均显示 AMLI 方法对预测结果有显著提升作用。我们还在 AMLI 的思想基础上结合聚类方法提出了一种基于类与类之间线性插值的 AMLI plus 方法，并且对 AMLI 与 AMLI plus 方法的有效性给出了理论证明。

Key words: multi-path ; linear interpolation ; sample optimization ; predicted effects

一、INTRODUCTION

日常通过机器学习模型进行预测时，我们经常会面临样本量不足，部分数据缺失或者观测误差较大等一系列问题，特别是针对小样本数据集而言，如何通过有效的样本优化技术提升模型的预测效果是非常具有研究意义的。

对于数值型数据量增加的方法最早可以追溯到插值法，De Boor(1978)提出三次样条插值，Mitas and Mitasova(1999)提出的空间插值法，Lu and Wong (2008)提出一种自适应反距离空间插值技术利用邻域之间的距离与权重成反比来进行插值. Efron (1992)提出在 jackknife 基础上使用 Bootstrap 重抽样法;Chawla et al.(2002)提出 SMOTE 过采样法;Pan and Yang (2009) 提出的迁移学习方法将不同类型的标签样本同时建模，丰富模型训练的样本量. Fernandez (2018)提出的 SMOTE 平滑法，SMOTE 平滑法通过随机选择一个样本，从 K 近邻中随机选择多个样本构建出新样本。

随着大数据时代的到来，机器学习、深度学习领域也有一些样本优化方面的研究。Zhu (2005) 提出的主动学习、半监督学习使用原样本空间已有的样本，用一定的算法为无标签样本打上高质量标签，达到样本优化的效果。Eisenberger et al.(2021)提出一种基于 NeuroMorph 网络的无监督形状插值法。Kokol et al.(2022)提出的综合数据学习法，证明了在统计机器学习的情况下，小样本可能比低质量的大样本更好。Zhou et al.(2022) 提出了一种新的改进的多尺度边缘标记图神经网络 (MEGNN)，通过获取尽可能多的特征信息来处理小样本问题。

在进行扩充样本量的相关处理时，往往也会面临对于增加的数据样本分布与实际样本分布(无法知晓)的偏差过大，并且我们日常观测到的数据经常是含有噪音的，针对有噪音的数

据进行样本扩充往往会更加加剧观测噪音对预测结果的影响．本文所提出的 AMLI 方法可以有效的解决这些问题，并且可以保证增加的样本大部分都是有效样本（这里的有效样本泛指与实际分布误差较小的样本）.AMLI 方法主要是在线性插值法的基础上对原始数据进行样本扩充，其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，将每个子空间各随机提取一个样本作为一个类，然后对于同类中的样本进行线性插值，既 K-path linear interpolation。AMLI 方法需要提前给定两个超参数 K 和 η，参数 K 的直观解释是各个特征子空间中存在的样本数量，η 是对样本进行线性插值的单位距离插入样本的个数。在后文的模拟以及实证研究中我们会发现，参数 K 的选取至关重要．针对不同的样本 K 的取值也不一样，通过选取合适的超参数，可以扩充大量的有效样本并且观测值与实际值误差较大的样本占比将会减小，调整了误差样本占比结构，达到了样本优化的目的，这也在很大程度上降低了观测噪音对预测结果的影响。

本文结构如下：第 2 节主要阐述了 AMLI 方法应该满足的假设以及具体的实施步骤；第 3 节通过 6 组蒙特卡洛模拟分析了 AMLI 方法在不同情况下的表现，探究了超参数的取值规律，并且与其他的插值方法进行对比；第 4 节结合了多种机器学习方法对模拟数据与实际数据进行预测，探究 AMLI 方法在预测方面的优化表现；第 5 节则是对于 AMLI 方法的有效性给予理论证明；第 6 节则是阐述了在 AMLI 方法基础上进行类与类之间线性插值的 AMLI plus 方法，并给出理论证明。

二、Research Hypothesis and Methodology Statement

本章节主要阐述 AMLI 方法步骤和使用 AMLI 方法应当满足的一些假设。

对于给定训练数据集：

$$T=\{(\mathbf{x}_1,y_1)(\mathbf{x}_2,y_2),...(\mathbf{x}_N,y_N)\},$$

其中，$\mathbf{x}_i=(x_i^{(1)},...,x_i^{(n)})\in\chi\subseteq R^n$ 为实例的特征向量，$y_i\in R$ 是相应的输出，$i=1,2,...,N,$ N 表示样本容量，n 为特征维度。我们假设：

$$y_i=f\left(\mathbf{x}_i-\varepsilon_i\right)+\tilde{\varepsilon}_i,\ \ i=1,2,\ldots,N,,$$

其中，$f\left(\cdot\right)$ 是连续函数，ε_i 是独立且同分布的观测噪音，$\tilde{\varepsilon}_i$ 为模型误差。AMLI 方法具体步骤如下：

首先，给定超参数 K，将特征空间 χ 划分为 N/K 个特征子空间，每个子空间均包含 K 个观测样本，从每个子空间随机取一个样本组成一个集合 $S_d,d=1,2,\ldots,k,$，则有：

$$T=\{S_1,S_{2,...,}S_K\},$$

其中，每个 S_d 包含的样本量为 N/K.

$\exists\mathbf{x}_0\in\chi,$对于 $\forall i,$有 $\mathbf{x}_0^{(i)}=\inf_{x\in花x}\{x^{(i)}\}$,我们将 $\mathbf{x}_0$ 称为特征空间最小点。设 $L(\cdot)$ 为特征空间 χ 的距离度量函数,$S_d=\{\mathbf{x}^{(d,1)},...\mathbf{x}^{(d,N/K)}\}$,$\mathbf{x}^{(d,1)}=\arg\min_{\mathbf{x}\in S_d}L(\mathbf{x},\mathbf{x}_0)$，我们将每个样本

类中距离 $\mathbf{x}_0$ 最近的点称为样本类最小点，其中 $d=1,...,K$。$\mathbf{x}^{(d,h+1)}=\underset{\{\mathbf{x}:\mathbf{x}\in S_d,\mathbf{x}\neq\mathbf{x}^{(d,1)},...,\mathbf{x}^{(d,h)}\}}{\arg\min} L(\mathbf{x},\mathbf{x}^{(d,h)})$ 为 S_d 中除 $\mathbf{x}^{(d,1)},...,\mathbf{x}^{(d,h)}$ 距离 $\mathbf{x}^{(d,h)}$ 最近的观测点，$h=1,2,\ldots,N/K-1$。

确定样本类最小点 $\mathbf{x}^{(d,1)}$ 后，搜索其所在集合所有样本找出在特征空间中距离 $\mathbf{x}^{(d,1)}$ 最近的点 $\mathbf{x}^{(d,2)}$，以及 $\mathbf{x}^{(d,3)},...,\mathbf{x}^{(d,N/K)}$。我们定义单位距离填充参数 η，我们在特征空间中使用线性插值法插入 $\sum_{i=1}^{N/K-1}\left\lfloor \eta\cdot L(\mathbf{x}^{(d,i)},\mathbf{x}^{(d,i+1)})\right\rfloor$（为了方便起见，本文对于插入的样本数量都是向下取整，后文不再特别标明）数量的虚拟样本，并且插入的样本之间是等距的。$\mathbf{x}_d^{(h,h+1,i)}$ 表示 S_d 中，$\mathbf{x}^{(d,h)}$ 与 $\mathbf{x}^{(d,h+1)}$ 间插入的第 i 个虚拟样本，$i=1,...,\eta\cdot L(\mathbf{x}^{(d,h)},\mathbf{x}^{(d,h+1)})$,对于 $\mathbf{x}_d^{(h,h+1,i)}$ 及其对应的输出 $y_d^{(h,h+1,i)}$ 满足：

$$\mathbf{x}_d^{(h,h+1,i)}=\mathbf{x}^{(d,h)}+i\cdot\frac{\mathbf{x}^{(d,h+1)}-\mathbf{x}^{(d,h)}}{\eta\cdot L(\mathbf{x}^{(d,h)},\mathbf{x}^{(d,h+1)})+1},$$
$$y_d^{(h,h+1,i)}=y^{(d,h)}+i\cdot\frac{y^{(d,h+1)}-y^{(d,h)}}{\eta\cdot L(\mathbf{x}^{(d,h)},\mathbf{x}^{(d,h+1)})+1}.$$

对于所有的样本类进行上述线性插值，插值数量为 $\sum_{d=1}^{K}\sum_{i=1}^{N/K-1}\eta\cdot L(\mathbf{x}^{(d,i)},\mathbf{x}^{(d,i+1)})$，并将所有的虚拟样本添加到数据集中。

Algorithm1 AMLI 方法

Require：参数 K，η

距离度量方法 $L(\cdot)$，eg: Euclidean distance 特征空间最小点 $\mathbf{x}_0$

集合 $S_1,S_{2,...,}S_K$

For d = 1,⋯,K do

$\mathbf{x}^{(d,1)}=\underset{\mathbf{x}\in S_d}{\arg\min} L(\mathbf{x},\mathbf{x}_0)$（确定样本类最小点）

For $h=1,\ldots,N/K-1$ do

$\mathbf{x}^{(d,h+1)}=\underset{\{\mathbf{x}:\mathbf{x}\in S_d,\mathbf{x}\neq\mathbf{x}^{(d,1)},...,\mathbf{x}^{(d,h)}\}}{\arg\min} L(\mathbf{x},\mathbf{x}^{(d,h)})$

For i= 1,2,⋯, $\eta\cdot L(\mathbf{x}^{(d,h)},\mathbf{x}^{(d,h+1)})$ do

$$\mathbf{x}_d^{(h,h+1,i)} = \mathbf{x}^{(d,h)} + i \cdot \frac{\mathbf{x}^{(d,h+1)} - \mathbf{x}^{(d,h)}}{\eta \cdot L(\mathbf{x}^{(d,h)}, \mathbf{x}^{(d,h+1)}) + 1},$$(确定插入样本特征)

$$y_d^{(h,h+1,i)} = y^{(d,h)} + i \cdot \frac{y^{(d,h+1)} - y^{(d,h)}}{\eta \cdot L(\mathbf{x}^{(d,h)}, \mathbf{x}^{(d,h+1)}) + 1}.$$（确定插入样本的输出）

添加（$\mathbf{x}_d^{(h,h+1,i)}$， $y_d^{(h,h+1,i)}$）到数据集 T 中

end

end

end

AMLI 方法要将原本的特征空间划分为具有等量样本的 N/K 个子空间，然后在每个子空间中随机选取一个样本作为一个样本类，而在实际应用中进行样本分类的操作时，可以对数据集所有观测样本计算与特征空间最小点的距离，从小到大依次归类到各个样本类中。

三、simulation experiments

3.1 Monte Carlo simulations

本节将通过 6 组蒙特卡洛模拟探究通过 AMLI 方法增加的样本对于样本整体的优化效果。为简单起见以及达到更好的可视化效果，定义特征维度 n=1，选用的真实函数关系为 $y=x^3$,由于样本的观测值与真实值之间存在一定的误差，为了模拟这一效果，我们数据生成后，对样本添加噪音。

对于数据的分布生成以及噪音和样本量的设定如下：

模拟 1：N=200，x~U（-2.5，2.5），ε ~N（0，1）

模拟 2：N=500，x~U（-2.5，2.5），ε ~N（0，1）

模拟 3：N=800，x~U（-2.5，2.5），ε ~N（0，1）

模拟 4：N=200，x~N（0，2.5），ε ~ N（0，1）

模拟 5：N=200，x~t (5)，ε ~ N（0，1）

模拟 6：N=200，x~U（-2.5，2.5），ε ~ U（-1.732，1.732）

将模拟 1 作为对照组，模拟 2、3 为不同样本量的实验组；模拟 4、5 为不同特征分布的实验组；模拟 6 为不同噪音的实验组（噪音分布参数的选择目的是为了统一噪音的方差并确保期望为 0）。选取一些验证指标来检验 AMLI 方法对于原始样本进行处理后的优化效果，选取验证指标分别为误差大于 0.5、1、1.5、2、2.5 的样本占比，以及样本观测值与实际值之间的均方误差（MSE）：

$$p_{(\alpha)} = \frac{\sum_{i=1}^{N+\sum_{d=1}^{K}\sum_{i=1}^{N/K-1}\eta \cdot L(\mathbf{x}^{(d,i)}, \mathbf{x}^{(d,i+1)})} I\left(\left|\mathbf{x}_i - \mathbf{x}_i^*\right| > \alpha\right)}{N + \sum_{d=1}^{K}\sum_{i=1}^{N/K-1}\eta \cdot L(\mathbf{x}^{(d,i)}, \mathbf{x}^{(d,i+1)})},$$

$$MSE(\mathbf{x},\mathbf{x}^*)=\frac{\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{x}_i^*)^2}{N},$$

其中，$\mathbf{x}$ 为样本观测值，$\mathbf{x}^*$ 为样本真实值，$N+\sum_{d=1}^{K}\sum_{i=1}^{N/K-1}\eta\cdot L(\mathbf{x}^{(d,i)},\mathbf{x}^{(d,i+1)})$ 为 AMLI 优化后的样本总量，α =0.5，1，1.5，2，2.5。由于每次的实验具有随机性，我们对于每个模拟选取不同的超参数多次调试，每进行 100 次实验分别计算验证指标的平均值。

为了更加直观感受到 AMLI 方法对于样本优化的效果，本文首先详细阐述模拟 1 的 AMLI 方法处理过程。我们在(-2.5，2.5)的区间内均匀生成 200 个样本（见图 1.1）；对特征变量添加服从标准正态分布的噪音（见图 1.2）；为了达到更好的可视化效果，我们仅将原始样本分为 4 类（仅仅只是为了达到更好可视化效果，验证指标并不是最优的），既令超参数 K=4，分别用不同颜色标注（见图 1.3）；由于定义域的区间较小，我们可以将单位距离填充参数取值大一些，设置超参数 η =100 进行样本填充，经过填充后，样本量达到了 3035 个（见图 1.4）。

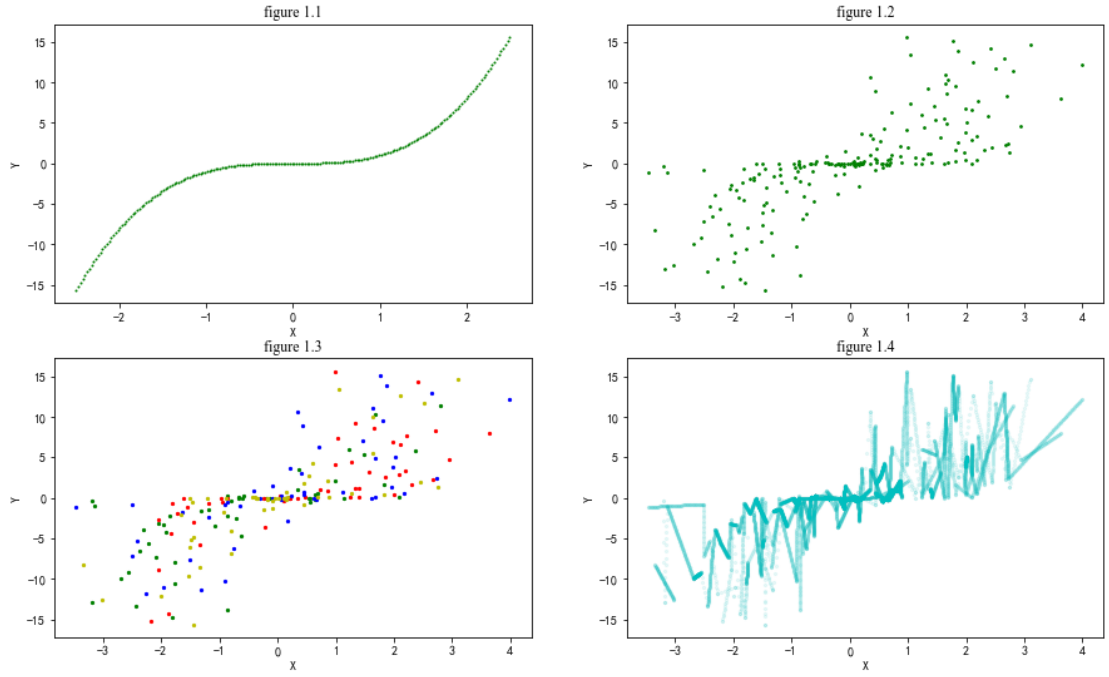


图 1　AMLI 方法处理过程

通过图 1.2 与图 1.4 的可视化效果对比来看，经过 AMLI 方法处理后，样本可以自适应的拟合出 x 与 y 之间的函数关系。

通过上述方法，对六个模拟进行参数调优，并计算验证指标，如下表所示：

表 1 蒙特卡洛模拟结果

模拟	MSE		$p_{(0.5)}$		$p_{(1)}$		$p_{(1.5)}$		$p_{(2)}$		$p_{(2.5)}$	
	处理前	处理后	处理前	处理后	处理前	处理后	处理前	处理后	处理前	处理后	处理前	处理后
1	0.957	0.762	0.604	0.570	0.316	0.251	0.130	0.083	0.036	0.022	0.010	0.006

2	0.980	0.774	0.614	0.567	0.313	0.258	0.132	0.089	0.043	0.023	0.011	0.005
3	0.977	0.789	0.619	0.571	0.308	0.254	0.131	0.090	0.043	0.025	0.011	0.005
4	0.99	0.701	0.609	0.549	0.317	0.232	0.133	0.073	0.044	0.017	0.011	0.003
5	0.982	0.756	0.611	0.552	0.306	0.245	0.127	0.085	0.038	0.024	0.010	0.005
6	0.987	0.742	0.706	0.632	0.411	0.297	0.126	0.057	0.000	0.000	0.000	0.000

Remark 1: 各模拟最优效果的选取是根据网格搜索法遍历所有超参数取值选 AMLI 处理后 MSE 最低的情况。

不难看出，经过 AMLI 方法处理后，样本的 MSE 以及各类误差样本的占比都得到了优化，添加的虚拟样本大部分都是有效样本，并且样本量的增加并不会显著弱化 AMLI 方法的优化效果；服从正态分布的样本具有更好的优化表现；即使是均匀分布的噪音，依然表现出良好的鲁棒性。在多种情况下，AMLI 方法对于数据的优化效果都有良好表现。

3.2 Analysis of hyperparameter taking values

本节探究对于 AMLI 方法中 K、η 参数的取值规律。首先探究参数 K 的取值规律，针对上述模拟，我们令参数 η =100，对 K=1，2…，200 进行遍历循环，并且每个取值进行一百次重复计算对指标取均值。AMLI 方法处理前后 MSE 指标变化趋势见下图：

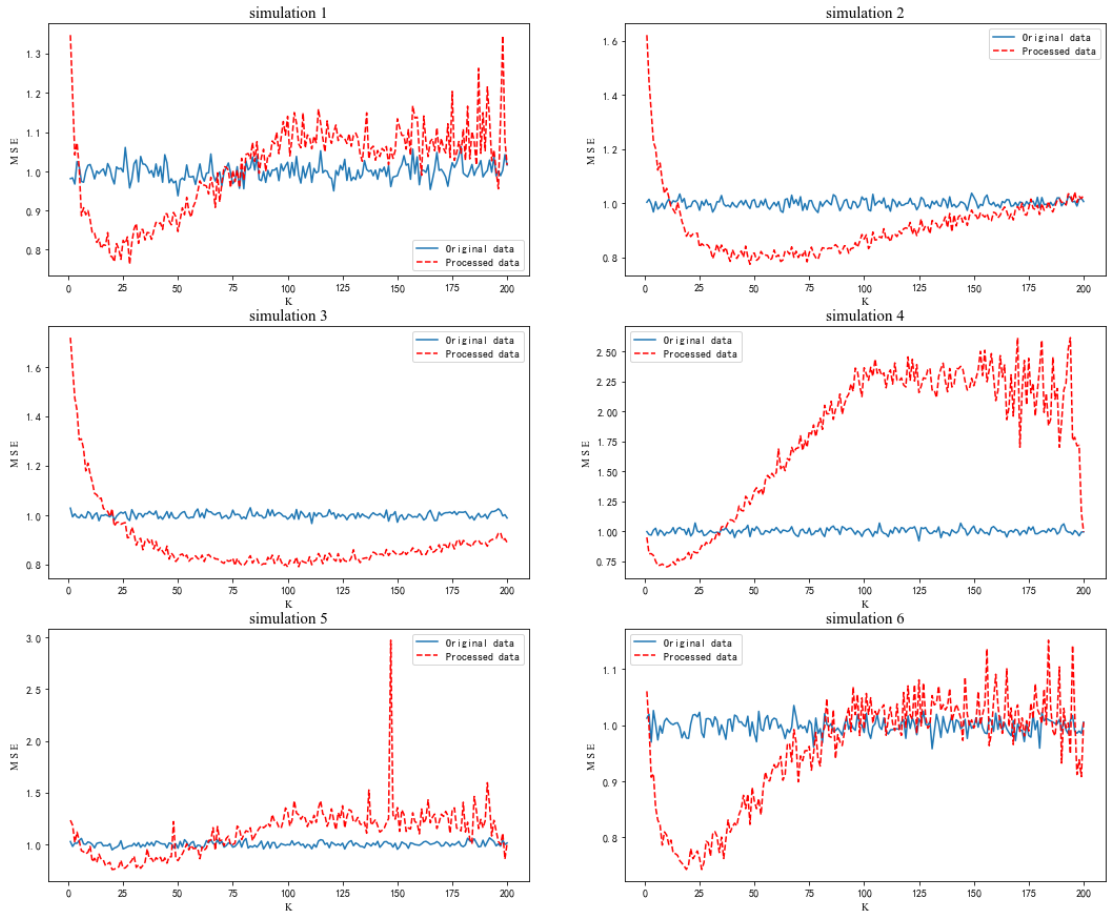
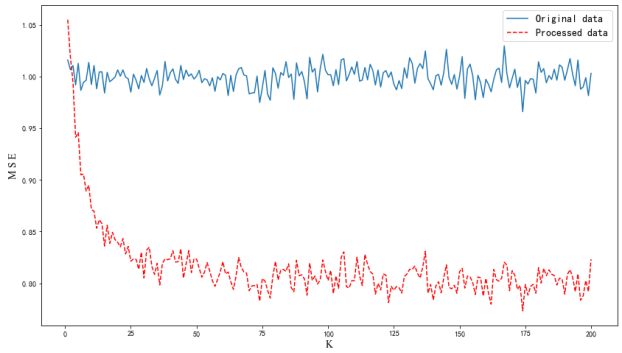


图 2 参数 K 不同取值下 MSE 变化趋势

在参数 η 取值固定的情况下，K 的最优取值会随着样本量的增加而增大；变量在不同分布下，k 的取值范围也会发生变化；噪音分布的变化对于 K 的最优取值影响不大。

接下来探究 η 参数的取值规律，我们针对模拟 1 的情况，在 K=21 取最优值条件下，令

参数 η =1，2，…，200 进行遍历循环，结果见图 3。不难发现，使用 AMLI 方法优化后的样本 MSE 值，随着填充参数 η 的增加而波动减小。



图三 参数 η 不同取值下 MSE 变化趋势

3.3 与其它插值方法对比

通过以上实验可发现 AMLI 方法可以极大扩充有效样本、减小样本整体的均匀误差，使得误差较大的样本占比较小。AMLI 方法主要是基于插值的思想进行样本优化，对于插值方法的研究主要有线性插值、二次样条插值以及三次样条插值，本节主要将 AMLI 方法与这些方法进行对比，体现 AMLI 方法的优越性。

依旧选取模拟 1 为对照组，模拟 2-6 为实验组。运用不同的插值方法进行处理，插值数量固定在 3500-4000 之间，仅选取处理后样本与真实值的均方误差作为评价指标，针对不同的模拟进行一百次实验并取均值，结果见表 2。不难看出，AMLI 方法处理后的样本 MSE 相比其他方法而言，达到了十分显著的优化效果。

表 2 不同插值方法下的样本 MSE

模拟	MSE			
	AMLI 方法	线性插值	二次样条插值	三次样条插值
1	0.762	1.385	5.542	9.452
2	0.774	1.618	6.434	8.976
3	0.789	1.812	6.441	8.338
4	0.701	1.578	18.156	42.997
5	0.756	1.203	10.322	29.542
6	0.742	0.959	4.802	6.619

四、Application of AMLI Method in Machine Learning

本节主要探究 AMLI 方法结合机器学习模型在数据预测方面的表现，主要分为模拟数据预测与实际数据预测两部分。使用旁置法，按照 7：3 的比例划分训练集与测试集，对于训练集我们用 AMLI 方法进行样本优化；机器学习方法我们选取 K 近邻法（KNN）、前馈神经网络（FNN）、梯度提升决策树（GBDT）以及随机森林（RF）；选择 MSE 作为损失函数；各模型超参数以及 AMLI 方法参数 K、η 均进行多次调试，取最优值。距离函数选择 Euclidean distance：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})^2$$

$$\mathrm{L}\left(x_i, x_j\right)=\left(\sum_{l=1}^{n}\left(x_i^l-x_j^l\right)^2\right)^{\frac{1}{2}}$$

4.1 Simulated data prediction

我们考虑模拟生成的数据特征维度 n=3，样本量为 1000 个的模拟样本，不同特征维度服从不同的分布，$\mathbf{x}^1 \sim N(0,3), \mathbf{x}^2 \sim U(-3,3), \mathbf{x}^3 \sim t(5),$ 随机生成 $\mathbf{w}_{(3,2)}^1, \mathbf{w}_{(2,1)}^2$ 的权重向量，令 $\mathbf{Y}=\mathbf{X} \cdot \mathbf{w}_{(3,2)}^1 \cdot \mathbf{w}_{(2,1)}^2+\tilde{\varepsilon}$,生成数据后对 $\mathbf{X}$ 添加服从高斯分布的噪音，然后划分测试集与训练集，对训练集进行 AMLI 方法处理，令 K=40，η =5，处理后的样本总量为 9810 个。

通过表 3 可以发现，经过 AMLI 方法处理过后的数据，训练得到的模型在预测方面的 MSE 更小，预测结果更加准确。

表 3 模拟数据预测结果

MSE	KNN	FNN	GBDT	RF
原 MSE	1.70	0.942	1.210	1.507
处理后 MSE	1.07	0.713	1.008	1.320

4.2 Actual data prediction

本节主要考虑共享单车租赁情况的预测数据，我们对某城市共享单车租赁需求进行预测，该数据集包含了 season、holiday、temp、registered 等多个变量（见表 4）。我们用 AMLI 方法对实际数据进行样本优化，一方面结合机器学习方法探究 AMLI 方法在实际预测活动中的优化表现；另一方面，由于该数据集含有多个分类数据，我们可以探究违背 AMLI 方法假设的情况发生时，AMLI 方法在预测方面是否能达到很好的优化效果。

表 4 变量指标说明

变量名称	变量定义
season	1=春天
	2=夏天
	3=秋天
	4=冬天
holiday	1=节假日 0=非节假日
working day	1=工作日
	0=周末
weather	1：晴天，多云
	2：雾天，阴天
	3：小雪，小雨
	4：大雨，大雪，大雾
temp	气温摄氏度

atemp	体感温度
humidity	湿度
windspeed	风速
casual	非注册用户个数
registered	注册用户个数
count	总租车人数

该数据集共含有 7620 个观测样本，我们分别选取 1000、3000、7620 个样本探究 AMLI 方法在样本量不足、样本量一般以及样本量充足情况下结合机器学习方法的预测优化情况。

表 5 实际数据预测结果

样本量	超参数	处理后样本量	MSE	KNN	FNN	GBDT	RF
1000	K=25 η =5	191057	处理前	208.23	0.9677	88.152	231.21
			处理后	46.34	0.0791	68.394	158.94
3000	K=40 η =3	290856	处理前	42.757	0.3342	22.794	43.679
			处理后	26.020	0.0084	14.301	25.451
7620	K=65 η =1	428277	处理前	17.464	0.1168	9.879	23.521
			处理后	6.798	0.0051	6.981	2.472

通过表 5 可知，在不同数据量的情况下，AMLI 方法对于预测效果都达到了一定优化。

五、proof

本节主要证明对于满足 AMLI 假设的样本，经过 AMLI 方法处理后，为什么样本平均观测误差减小以及不同误差的样本占比得到调整。

AMLI 的思想是将原本的特征空间划分为具有 K 个样本的 N/K 个子空间，将每个子空间各随机提取一个样本作为一个类，则原始数据集可划分为 K 类，然后对于同类中距离相近的两个样本之间进行线性插值，我们可以理解为 AMLI 是对距离相近的两个子空间之间通过 K 条线路进行线性插值的方法。

对于数据集 $T=\{(x_1,y_1)(x_2,y_2),...(x_N,y_N)\}$ 我们选取假设空间中存在等量样本的两个子空间 $\chi^{(1)},\chi^{(2)}\subseteq\chi\subseteq R$，且 $\chi^{(1)}\cap\chi^{(2)}=\varnothing$，在这两个子空间中，我们假设存在以下共同关系：

$$y=f(x^*)+\tilde{\varepsilon}=f(x-\varepsilon)+\tilde{\varepsilon}, \tag{1}$$

其中，x^* 为 x 剔除观测噪音后的真实值，$x=x^*+\varepsilon$ ，ε 为噪音项且 $\varepsilon\sim N(0,\sigma^2)$，$\tilde{\varepsilon}_i$ 为模型误差。

根据我们的假设，$f(\cdot)$ 是一个连续的函数，若 $\chi^{(1)},\chi^{(2)}\to 0$，且 $dist(\chi^{(1)},\chi^{(2)})\to 0$，$\chi^{(1)}\cap\chi^{(2)}=\varnothing$，对于 $f(\cdot)$ 可近似看作一个线性函数 $g(\cdot)$,则对于（1）可转化为：

$$y = g(x^*) + \varepsilon_1 + \tilde{\varepsilon} = g(x - \varepsilon) + \varepsilon_1 + \tilde{\varepsilon} \tag{2}$$

其中，ε_1 为线性拟合误差项且 $\varepsilon_1 \to 0$。

由于我们选取的 $\chi^{(1)}, \chi^{(2)}$ 中的样本是等量的，因此假设在两个子空间中各存在 K 个观测样本，即 $\mathbf{x}^{(j)} = (x_1^{(j)}, ..., x_k^{(j)}) = (x^{*(j)}_1 + \varepsilon_1^{(j)}, ..., x^{*(j)}_k + \varepsilon_k^{(j)}) \in \chi^{(j)}, j = 1, 2.$ 此时在空间 $\chi^{(1)}, \chi^{(2)}$ 中均匀噪音绝对值的期望为：

$$\begin{aligned} E(\frac{\sum_{j=1}^{2}\sum_{i=1}^{k}\left|\varepsilon_i^{(j)}\right|}{2k}) &= \int_{-\infty}^{+\infty} |t| \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{t^2}{2\sigma^2}} dt \\ &= \sqrt{\frac{2}{\pi}}\cdot\sigma \int_0^{+\infty} \frac{t}{\sigma} t^{-\frac{t^2}{2\sigma^2}} d\frac{t}{\sigma} \\ &= \sqrt{\frac{2}{\pi}}\cdot\sigma \int_0^{+\infty} e^{-\frac{t^2}{2\sigma^2}} d\frac{(\frac{t}{\sigma})^2}{2} \\ &= \sqrt{\frac{2}{\pi}}\cdot\sigma\cdot\Gamma(1) = \sqrt{\frac{2}{\pi}}\cdot\sigma \end{aligned}$$

噪音大于 0.5 的观测样本占比的期望为：

$$\begin{aligned} E(\frac{\sum_{j=1}^{2}\sum_{i=1}^{k} I(\left|x_i^{(j)} - x^{*(j)}_i\right| > 0.5)}{2k}) &= E(\frac{\sum_{j=1}^{2}\sum_{i=1}^{k} I(\left|\varepsilon_i^{(j)}\right| > 0.5)}{2k}) \\ &= \frac{\sum_{j=1}^{2}\sum_{i=1}^{k} (P(\varepsilon_i^{(j)} > 0.5) + P(\varepsilon_i^{(j)} < -0.5))}{2k} \\ &= \frac{\sum_{j=1}^{2}\sum_{i=1}^{k} (P(\frac{\varepsilon_i^{(j)}}{\sigma} > \frac{0.5}{\sigma}) + P(\frac{\varepsilon_i^{(j)}}{\sigma} < \frac{-0.5}{\sigma}))}{2k} \\ &= \frac{\sum_{j=1}^{2}\sum_{i=1}^{k} \Phi(\frac{-0.5}{\sigma})}{k} = 2\Phi(\frac{-0.5}{\sigma}) \end{aligned}$$

我们从子空间中任选两个样本进行线性插值，重复 K 次。AMLI 方法是根据两个样本之间的距离确定的插值样本的数量，为了简单方便起见，我们假设每次插入的样本量为 m，其余情况类似可证明，第 i 次插值的样本为：

$$\mathbf{x}'_i=(x'_{i,1},x'_{i,2},...,x'_{i,m})$$
$$=(x_i^{(1)}+\frac{x_{i,}^{(2)}-x_i^{(1)}}{m+1},x_i^{(1)}+2\cdot\frac{x_{i,}^{(2)}-x_i^{(1)}}{m+1},...,x_i^{(1)}+m\cdot\frac{x_{i,}^{(2)}-x_i^{(1)}}{m+1})$$
$$=(x^{*(1)}_i+\varepsilon_i^{(1)}+\frac{x^{*(2)}_i-x^{*(1)}_i+\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1},...,x^{*(1)}_i+\varepsilon_i^{(1)}+m\cdot\frac{x^{*(2)}_i-x^{*(1)}_i+\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1}),$$

其中，i=1,⋯,k，对应的输出为 $y'_{i,d}=y_i^{(1)}+d\cdot\frac{y_i^{(2)}-y_i^{(1)}}{m+1}$，根据公式（2）易得 $x'_{i,d}$ 的噪音 $\varepsilon'_{i,d}=\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1}$，d=1,⋯,m.

进行 k 次插值后，均匀噪音期望为：

$$E(\frac{\sum_{j=1}^{2}\sum_{i=1}^{k}\left|\varepsilon_i^{(j)}\right|+\sum_{i=1}^{k}\sum_{d=1}^{m}\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|}{2k+k\cdot m})$$
$$=\frac{\sum_{j=1}^{2}\sum_{i=1}^{k}E(\left|\varepsilon_i^{(j)}\right|)+\sum_{i=1}^{k}\sum_{d=1}^{m}E(\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|)}{2k+k\cdot m}$$
$$=\frac{\sum_{j=1}^{2}\sum_{i=1}^{k}E(\left|\varepsilon_i^{(j)}\right|)+\frac{1}{m+1}\sum_{i=1}^{k}\sum_{d=1}^{m}E(\left|(m-d+1)\varepsilon_i^{(1)}+d\varepsilon_i^{(2)}\right|)}{2k+k\cdot m}$$
$$=\frac{2+\frac{1}{m+1}\sum_{d=1}^{m}\sqrt{(m-d+1)^2+d^2}}{2+m}\cdot\sigma\sqrt{\frac{2}{\pi}}$$
$$<\frac{2+\frac{1}{m+1}\sum_{d=1}^{m}\sqrt{(m-d+1)^2+d^2+2d(m-d+1)}}{2+m}\cdot\sigma\sqrt{\frac{2}{\pi}}$$
$$=\frac{2+\frac{1}{m+1}\sum_{d=1}^{m}\sqrt{(m+1)^2}}{2+m}\cdot\sigma\sqrt{\frac{2}{\pi}}=\sigma\sqrt{\frac{2}{\pi}}$$

可得，经过 AMLI 方法优化后得样本均匀噪音减小。噪音大于 0.5 的样本占比的期望为:

$$E\left(\frac{\sum_{j=1}^{2}\sum_{i=1}^{k} I\left(\left|\varepsilon_i^{(j)}\right|>0.5\right)+\sum_{i=1}^{k}\sum_{d=1}^{m} I\left(\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|>0.5\right)}{2k+k\cdot m}\right)$$

$$=\frac{4k\cdot\Phi\left(\frac{-0.5}{\sigma}\right)+\sum_{i=1}^{k}\sum_{d=1}^{m} p\left(\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_i^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|>0.5\right)}{2k+k\cdot m}$$

$$=\frac{4k\cdot\Phi\left(\frac{-0.5}{\sigma}\right)+\sum_{i=1}^{k}\sum_{d=1}^{m} p\left(\frac{\left|(m-d+1)\varepsilon_i^{(1)}+d\cdot\varepsilon_i^{(2)}\right|}{\sqrt{(m-d+1)^2+d^2}\cdot\sigma}>\frac{0.5(m+1)}{\sqrt{(m-d+1)^2+d^2}\cdot\sigma}\right)}{2k+k\cdot m}$$

$$=\frac{4\Phi\left(\frac{-0.5}{\sigma}\right)+2\sum_{d=1}^{m}\Phi\left(-\frac{0.5(m+1)}{\sqrt{(m-d+1)^2+d^2}\cdot\sigma}\right)}{2+m}$$

$$<\frac{4\Phi\left(\frac{-0.5}{\sigma}\right)+2\sum_{d=1}^{m}\Phi\left(-\frac{0.5(m+1)}{\sqrt{(m-d+1)^2+d^2+2d(m-d+1)}\cdot\sigma}\right)}{2+m}=2\Phi\left(\frac{-0.5}{\sigma}\right)$$

经过 AMLI 方法处理后的数据，误差大于 0.5 的样本占比减小。

六、EXTENSIONS

6.1 AMLI plus

通过上文的模拟实验可以发现，在 AMLI 算法中参数 K 的选取至关重要，参数 K 的最优取值涉及到大量的调参计算，并且受到随机性的影响我们很难保证选取的 K 值在任何时候的表现都是最优的。在 AMLI 方法难以达到很好的效果时，可以考虑换一种插值的思路，结合聚类方法对样本进行类与类之间的线性插值，这种方法我们称之为 AMLI plus。下面我们详细阐述下 AMLI plus 方法的基本步骤。

首先，我们根据样本的分布情况对所有观测样本进行聚类，假设聚类个数为 K，将假设空间划分为 K 个子空间，每个子空间中包含同一个类别的所有观测样本，即：

$$T=\{(\mathbf{x}_1,y_1)(\mathbf{x}_2,y_2),...(\mathbf{x}_K,y_K)\}=\{G_1,G_2,...,G_K\},$$

其中，G_1 表示其类的中心距离特征空间最小点$\mathbf{x}_0$ 最近的类，类的中心$\bar{\mathbf{x}}^{(d)}=\frac{1}{n^{(d)}}\sum_{i=1}^{n^{(d)}}\mathbf{x}_i^{(d)}$，d=1,⋯,K，$n^{(d)}$为$G_d$中样本个数，$G_{d+1}$满足$\bar{\mathbf{x}}^{(d+1)}=\underset{\{\bar{\mathbf{x}}:\bar{\mathbf{x}}\neq\bar{\mathbf{x}}^{(1)},...,\bar{\mathbf{x}}^{(d)}\}}{\arg\min}\left(L(\bar{\mathbf{x}}^{(d)},\bar{\mathbf{x}})\right)$，聚类方法的选择具有多样性，可以选用 K-means、x-means 或 Ester et al.（1996）提出的可以剔除噪声点的 DBSCAN 方法等等；进行类与类之间的插值，定义单位距离填充参数η，对所有的$\mathbf{x}_i^{(d)}\in G_d$，使其分别与$G_{d+1}$中所有样本之间进行$n^{(d)}\cdot n^{(d+1)}$次线性插值，插值样本个数为$\sum_{j=1}^{n^{(d+1)}}\sum_{i=1}^{n^{(d)}}\eta\cdot L(x_i^{(d)},x_j^{(d+1)})$，并且插值的样本之间也是等距的。

Algorithm2 AMLI plus 方法

Require:单位距离填充参数 η

Require: 距离度量方法 $L(\cdot)$

Require: 特征空间最小点 $\mathbf{x}_0$

集合 $G_1, G_2, ..., G_K$

For d=1,⋯,K-1 do

For h = 1,⋯, $n^{(d)}$ do

For j = 1,⋯, $n^{(d+1)}$ do

For i = 1,⋯, $\eta \cdot L(\mathbf{x}^{(d+1,j)}, \mathbf{x}^{(d,h)})$ do

$$\mathbf{x}_d^{(h,j,i)} = \mathbf{x}^{(d,h)} + i \cdot \frac{\mathbf{x}^{(d+1,j)} - \mathbf{x}^{(d,h)}}{\eta \cdot L(\mathbf{x}^{(d+1,j)}, \mathbf{x}^{(d,h)}) + 1}, \text{(确定插入样本的特征)}$$

$$y_d^{(h,j,i)} = y^{(d,h)} + i \cdot \frac{y^{(d+1,j)} - y^{(d,h)}}{\eta \cdot L(y^{(d+1,j)}, y^{(d,h)}) + 1}, \text{（确定插入样本的输出）}$$

添加（$\mathbf{x}_d^{(h,j,i)}$， $y_d^{(h,j,i)}$）到数据集 T 中

end

end

end

end

return T (Resulting dataset)

6.2 the proof of AMLI plus

本节主要针对 AMLI plus 方法的有效性给予证明，证明思路与 AMLI 基本相同，我们主要针对不同点进行论证。

与 AMLI 方法不同的是，DUK plus 方法是根据聚类个数，将假设空间划分为 K 个子空间，每个子空间中包含同一个类别的所有观测样本，因此各个子空间存在的样本量可能也不一样。假设对于两个相近的两个子空间 $\chi^{(1)}, \chi^{(2)}$ 中存在的样本量分别为 $n^{(1)}, n^{(2)}$，进行 $n^{(1)} \cdot n^{(2)}$ 次线性插值，为了简单起见，假设每次插入的样本数量为 m，插值后的均匀噪音为：

$$E(\frac{\sum_{i=1}^{n^{(1)}}\left|\varepsilon_i^{(1)}\right|+\sum_{i=1}^{n^{(2)}}\left|\varepsilon_i^{(2)}\right|+\sum_{i=1}^{n^{(1)}}\sum_{j=1}^{n^{(2)}}\sum_{d=1}^{m}\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_j^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m})$$

$$=\frac{\sum_{i=1}^{n^{(1)}}E(\left|\varepsilon_i^{(1)}\right|)+\sum_{i=1}^{n^{(2)}}E(\left|\varepsilon_i^{(2)}\right|)+\sum_{i=1}^{n^{(1)}}\sum_{j=1}^{n^{(2)}}\sum_{d=1}^{m}E(\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_j^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|)}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}$$

$$=\frac{n^{(1)}\sigma\sqrt{\frac{2}{\pi}}+n^{(2)}\sigma\sqrt{\frac{2}{\pi}}+\frac{1}{m+1}\sum_{i=1}^{n^{(1)}}\sum_{j=1}^{n^{(2)}}\sum_{d=1}^{m}E(\left|(m-d+1)\varepsilon_i^{(1)}+d\varepsilon_i^{(2)}\right|)}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}$$

$$=\frac{n^{(1)}+n^{(2)}+\frac{n^{(1)}n^{(2)}}{m+1}\sum_{d=1}^{m}\sqrt{(m-d+1)^2+d^2}}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}\cdot\sigma\sqrt{\frac{2}{\pi}}$$

$$<\frac{n^{(1)}+n^{(2)}+\frac{n^{(1)}n^{(2)}}{m+1}\sum_{d=1}^{m}\sqrt{(m-d+1)^2+d^2+2d(m-d+1)}}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}\cdot\sigma\sqrt{\frac{2}{\pi}}$$

$$=\frac{n^{(1)}+n^{(2)}+\frac{n^{(1)}n^{(2)}}{m+1}\sum_{d=1}^{m}m+1}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}\cdot\sigma\sqrt{\frac{2}{\pi}}=\sigma\sqrt{\frac{2}{\pi}}$$

可得，经过 AMLI plus 方法优化后的均匀噪音减小。噪音误差大于 0.5 的样本占比的期望为:

$$E(\frac{\sum_{i=1}^{n^{(1)}}I(\left|\varepsilon_i^{(1)}\right|>0.5)+\sum_{i=1}^{n^{(2)}}I(\left|\varepsilon_i^{(2)}\right|>0.5)+\sum_{i=1}^{n^{(1)}}\sum_{j=1}^{n^{(2)}}\sum_{d=1}^{m}I(\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_j^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|>0.5)}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m})$$

$$=\frac{\sum_{i=1}^{n^{(1)}}P(\left|\varepsilon_i^{(1)}\right|>0.5)+\sum_{i=1}^{n^{(2)}}P(\left|\varepsilon_i^{(2)}\right|>0.5)+\sum_{i=1}^{n^{(1)}}\sum_{j=1}^{n^{(2)}}\sum_{d=1}^{m}P(\left|\varepsilon_i^{(1)}+d\cdot\frac{\varepsilon_j^{(2)}-\varepsilon_i^{(1)}}{m+1}\right|>0.5)}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}$$

$$=\frac{2(n^{(1)}+n^{(2)})\Phi(\frac{-0.5}{\sigma})+\sum_{i=1}^{n^{(1)}}\sum_{j=1}^{n^{(2)}}\sum_{d=1}^{m}p(\frac{\left|(m-d+1)\varepsilon_i^{(1)}+d\cdot\varepsilon_j^{(2)}\right|}{\sqrt{(m-d+1)^2+d^2}\cdot\sigma}>\frac{0.5(m+1)}{\sqrt{(m-d+1)^2+d^2}\cdot\sigma})}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}$$

$$=\frac{2(n^{(1)}+n^{(2)})\Phi(\frac{-0.5}{\sigma})+2n^{(1)}n^{(2)}\sum_{d=1}^{m}\Phi(-\frac{0.5(m+1)}{\sqrt{(m-d+1)^2+d^2}\cdot\sigma})}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}$$

$$<\frac{2(n^{(1)}+n^{(2)})\Phi(\frac{-0.5}{\sigma})+2n^{(1)}n^{(2)}\sum_{d=1}^{m}\Phi(-\frac{0.5(m+1)}{\sqrt{(m-d+1)^2+d^2+2d(m-d+1)}\cdot\sigma})}{n^{(1)}+n^{(2)}+n^{(1)}n^{(2)}m}=2\Phi(\frac{-0.5}{\sigma})$$

可得，经过 AMLI plus 方法处理后的数据，误差大于 0.5 的样本占比减小。

Remark 2 :AMLI plus 的证明思想是把我们聚类后类别不同的样本划分到不同的子空间，并假设相近的子空间存在共同的线性关系，聚类个数以及聚类方法的选择应当尽可能满足这一要求。

七、CONCLUSION

本文介绍了一种基于线性插值思想的多路径样本扩充方法，该方法主要解决预测中样本量不足或观测样本与实际分布误差较大的问题。AMLI 方法可以极大程度扩充有效样本，并且降低样本噪音的影响。该方法实现简单，可以适应多种情况，对于预测效果提升明显。最后，我们在 AMLI 方法上提出了另一种基于类与类之间的线性插值方法——AMLI plus，该方法也可以达到很好的优化效果。总的来说，我们发现 AMLI 方法是稳健且有效的，非常适合于机器学习领域中处理样本量不足、观测噪音较大等一系列问题。

参考文献

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.

Efron B. Bootstrap methods: another look at the jackknife[M]//Breakthroughs in statistics. Springer, New York, NY, 1992: 569-593.

Fernández A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of artificial intelligence research, 2018, 61: 863-905.

Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2009, 22(10): 1345-1359.

ZHU X J. Semi-supervised learning literature survey[R]. Madison：University of Wisconsin-Madison，2005.

Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//kdd. 1996, 96(34): 226-231.

De Boor C, De Boor C. A practical guide to splines[M]. New York: springer-verlag, 1978.

Bao W, Lai W S, Zhang X, et al. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(3): 933-948.

Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: A synthetic knowledge synthesis[J]. Science Progress, 2022, 105(1): 00368504211029777.

Zhou Y, Zhi G, Chen W, et al. A new tool wear condition monitoring method based on deep learning under small samples[J]. Measurement, 2022, 189: 110622.

Mitas L, Mitasova H. Spatial interpolation[J]. Geographical information systems: principles, techniques, management and applications, 1999, 1(2).

Lu G Y, Wong D W. An adaptive inverse-distance weighting spatial interpolation technique[J]. Computers & geosciences, 2008, 34(9): 1044-1055.

Eisenberger M，Novotny D，Kerchenbaum G，et al. Neuromorph：Unsupervised shape interpolation and correspondence in One Go[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021: 7473-7483.