**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。我们介绍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活动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，调整样本结构，降低样本的均匀噪音，以此来提升机器学习模型的预测效果。The hyper-parameters have intuitive interpre-tations and typically require little tuning.（该方法的超参数具有直观的解释，通常需要很少的调优）。我们针对多种情况做了大量的模拟实验并且结合机器学习方法探究了AMLI方法对预测结果的提升效果。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，并且针对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

在我们日常通过机器学习模型进行预测时，经常会面临样本量不足的情况，这就会导致训练的模型泛化能力较差、容易过拟合等一系列问题，因此针对小样本数据集，如何通过有效的数据扩充技术将原本数据集扩大是非常具有现实意义的。

对于增加数据样本量的方法较早的有Efron (1992)提出的在jackknife上使用Bootstrap再砍刀重抽样法； Chawla et al.(2002)提出的SMOTE过采样法；Pan and Yang (2009) 提出的迁移学习方法，它可将不同类型的标签样本同时建模，丰富模型训练的样本量；数据合成法通过少量可用的样本生成更多的样本，即从原始数据分布的角度来进行，生成和真实数据分布相似的数据，达到样本增强的目的,例如Fernandez (2018)提出的SMOTE平滑法主要应用在小型数据集上来获得新的样本，实现方式是随机选择一个样本，从K近邻中随机选择多个样本构建出新样本，对于数值型数据量增加的方法最早可以追溯到的插值法，它可以有效的增加数量，达到减少总体误差的效果，例如B样条插值，三次样条插值De Boor(1978),以及近年来提出的最近邻插值法，双线性插值，双三次插值法，以及Bao W et al(2019)提出的视频插值法。

随着大数据时代的到来，在机器学习、深度学习领域也有一些针对扩充样本方面的研究。Zhu X（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，达到样本量增加的效果；Taylor and Nitschke(2018)提出的基于深度学习的数据增强法，它通过变换原有的数据生成新的数据来扩大训练集，达到防止神经网络学习到不相干信息的效果，从根本上提升整体性能; 近年来，CV数据增强法，也得到广泛地应用，其中有监督的数据增强法包括几何变换法，颜色变换法，随机擦除法等。无监督学习的数据增强法包括，Goodfellow I (2014)提出的GAN生成对抗网络模型，在整个过程中，生成器生成的数据像接近于真实数据分布，而判别器识别不出真假数据，从而达到构建更多新样本的目的；Cubuk et al.(2018)提出的 AutoAugment和PBA（Population Based Augmentation）方法通过模型自主学习出适合当前任务的数据增强方法; 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；以及Zhou et al.(2022)所提出基于深度学习的GNN 通过获取尽可能多的特征信息来处理小样本问题。

但是在我们进行增加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都是含有噪音的，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与预测结果的影响。本文所提出的DYK方法可以很好的解决这些问题，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），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。DYK方法需要提前给定两个超参数K和，参数K的直观解释是各个特征子空间中存在的样本数量，是对样本进行线性插值的单位距离插值数量的参数。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，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：第二章主要阐述了DYK方法应该满足的一些假设以及具体的实施步骤；第三章通过六组蒙特卡洛模拟探究DYK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；第四章通过模拟数据与实际数据结合多种机器学习方法探究DYK方法在预测方面的优化表现；第五章则是对于DYK方法的有效性给予理论证明；第六章则是阐述了在DYK方法思想基础上进行类与类之间插值的DYK plus方法，并给予理论证明。

# 二、Research Hypothesis and Methodology Statement

本章节主要阐述具体的DYK方法步骤及使用DYK方法应当满足的一些假设。

对于给定训练数据集：



其中，为实例的特征向量，是相应的输出， N表示样本容量，n为特征维度。

我们假设， 是连续函数, 是独立且同分布的观测噪音。DYK方法具体步骤如下：

设为特征空间的距离度量函数，对于有。将所有样本随机、均匀且不重复的分到集合中，，其中为全部样本的集合，为集合Sd中除 距离最近的点，,为集合的元素个数。

我们将特征空间中满足****点条件的点称为特征空间最小点，我们将每个集合中距离点最近的点称为样本子集最小点，确定样本子集最小点后，搜索其所在集合所有样本找出在特征空间中距离最近的点 ，以及；我们定义单位距离填充参数，我们在特征空间中使用线性插值法插入（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。表示集合中，与间插入的第i个虚拟样本， ,对于及其对应的输出满足：



对于所有的集合进行上述线性插值，并将所有的虚拟样本添加到训练数据集中。

|  |
| --- |
| **Algorithm1** DYK方法 |
| Require:样本分类集合个数K以及单位距离填充参数  Require: 距离度量方法，eg: Euclidean distance  Require: 特征空间最小点P |
| 遍历循环所有样本，依次进行随机、均匀归类,归类个数为K  For d = 1,…,K do  （确定每个样本中的样本子集最小点）  For h= 1,…, do    使用线性插值法插入数量的等距的虚拟样本，并储存到训练集中  **end**  **end** |

**Remark：**在进行样本分类的操作时，要尽可能保证每个类中的样本在特征空间中的分布是均匀的。

# 三、simulation experiments

## 3.1 Monte Carlo simulations

本节将通过六组蒙特卡洛模拟探究通过DYK方法增加的样本对于样本整理的优化效果。为简单起见以及达到更好的可视化效果，定义特征维度为1，选用的真实函数关系为y=x3,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：

模拟一：N=200，x~U（-2.5，2.5），noise~N（0，1）

模拟二：N=500，x~U（-2.5，2.5），noise~N（0，1）

模拟三：N=800，x~U（-2.5，2.5），noise~N（0，1）

模拟四：N=200，x~N（0，2.5），noise~ N（0，1）

模拟五：N=200，x~t (5)，noise~ N（0，1）

模拟六：N=200，x~U（-2.5，2.5），noise~ U（-1.732，1.732）

将模拟一作为对照组，模拟二、三为不同样本量的实验组；模拟四、五为不同特征分布的实验组；模拟六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）。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）：



其中，N为样本量，为样本观测值，为样本真实值。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，每进行一百次实验分别计算验证指标的平均值。

为了更加直观感受到DYK方法对于样本优化的效果，本文首先详细阐述模拟一的DYK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四类（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=100进行样本填充，经过填充后，样本量达到了3035个（见图1.4）。

图片包含 散点图

描述已自动生成

图1 DYK方法处理过程

通过图1.4与图1.2的可视化效果对比来看，经过DYK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 | | Error greater than 0.5 | | Error greater than 1.0 | | Error greater than 1.5 | | Error greater than 2.0 | | Error greater than 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 |
| 一 | 0.957 | 0.762 | 0.604 | 0.570 | 0.316 | 0.251 | 0.130 | 0.083 | 0.036 | 0.022 | 0.010 | 0.006 |
| 二 | 0.980 | 0.774 | 0.614 | 0.567 | 0.313 | 0.258 | 0.132 | 0.089 | 0.043 | 0.023 | 0.011 | 0.005 |
| 三 | 0.977 | 0.789 | 0.619 | 0.571 | 0.308 | 0.254 | 0.131 | 0.090 | 0.043 | 0.025 | 0.011 | 0.005 |
| 四 | 0.99 | 0.701 | 0.609 | 0.549 | 0.317 | 0.232 | 0.133 | 0.073 | 0.044 | 0.017 | 0.011 | 0.003 |
| 五 | 0.982 | 0.756 | 0.611 | 0.552 | 0.306 | 0.245 | 0.127 | 0.085 | 0.038 | 0.024 | 0.010 | 0.005 |
| 六 | 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:**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DYK处理后MSE最低的情况。

不难看出，经过DYK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DYK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，DYK方法对于数据的增强效果都有良好表现。

## 3.2 Analysis of hyperparameter taking values

本节探究对于DYK方法中K、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=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。DYK方法处理前后MSE指标变化趋势见下图：图形用户界面

描述已自动生成

图二 参数K不同取值下MSE变化趋势

在参数取值固定的情况下，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。

接下来探究参数的取值规律，我们针对模拟一的情况，在K=21取最优值条件下，令参数=1，2，…，200进行遍历循环，结果见图三。不难发现，使用DYK方法优化后的样本MSE值，随着填充参数的增加而波动减小。

背景图案

中度可信度描述已自动生成

图三 参数不同取值下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DYK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。本节主要将DYK方法与线性插值、二次样条插值以及三次样条插值方法进行对比，体现DYK方法的优越性。

依旧选取模拟一为对照组，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，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，DYK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 | | | |
| DYK方法 | 线性插值 | 二次样条插值 | 三次样条插值 |
| 一 | 0.762 | 1.385 | 5.542 | 9.452 |
| 二 | 0.774 | 1.618 | 6.434 | 8.976 |
| 三 | 0.789 | 1.812 | 6.441 | 8.338 |
| 四 | 0.701 | 1.578 | 18.156 | 42.997 |
| 五 | 0.756 | 1.203 | 10.322 | 29.542 |
| 六 | 0.742 | 0.959 | 4.802 | 6.619 |

# 四、Application of DYK Method in Machine Learning

本章节主要探究DYK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DYK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DYK方法参数K、均进行多次调试，取最优值。距离函数选择Euclidean distance：



## 4.1 Simulated data prediction

我们考虑模拟生成的预测数据特征维度n=3，样本量为1000个的虚拟样本，不同特征维度服从不同的分布，随机生成 的权重向量，令 ,划分测试集与训练集后，为训练集添加服从高斯分布的噪音并进行DYK方法处理。

通过表3可以发现，经过DYK方法处理过后的训练集，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 | 15.70 | 9.742 | 12.21 | 17.507 |
| 处理后MSE | 11.07 | 8.813 | 9.718 | 11.320 |

## 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）。我们运用DYK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DYK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DYK方法假设的情况发生时，DYK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DYK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 | MSE | KNN | FNN | GBDT | RF |
| N=1000 | 原MSE | 208.23 | 0.9677 | 88.152 | 231.21 |
| 处理后MSE | 143.34 | 0.0791 | 68.394 | 158.94 |
| N=3000 | 原MSE | 42.757 | 0.3342 | 22.794 | 43.679 |
| 处理后MSE | 26.020 | 0.0084 | 14.301 | 25.451 |
| N=7620 | 原MSE | 17.464 | 0.1168 | 9.879 | 23.521 |
| 处理后MSE | 6.798 | 0.0051 | 6.981 | 2.472 |

通过表5可知，在不同数据量的情况下，DYK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DYK假设的样本，经过DYK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。

对于数据集我们选取假设空间中存在等量样本的两个子空间 ，且，在这两个子空间中，我们假设存在以下共同关系：

 （1）

其中，为 剔除观测噪音后的真实值， ，为噪音误差项且。

根据我们的假设，是一个连续的函数，若，且，，对于可近似看作一个线性函数 ,则对于（1）可转化为：

 （2）

其中，为线性拟合误差项且。

由于我们选取的中的样本是等量的，因此假设在两个子空间中各存在k个观测样本，即此时在空间中均匀噪音绝对值的期望为：



噪音大于0.5的观测样本占比的期望为：



我们从子空间中任选两个样本进行k次插值，每次插入m个样本:



其中，i=1,…,k，对应的输出为，根据公式（2）易得的噪音，d=1,…,m.

进行k次插值后，均匀噪音期望为：



易得，当m=1,2,…,时，有，故插值后的均匀噪音减小。从极限的角度来看，当 时，DYK方法可以使得均匀噪音减少至 。

噪音大于0.5的样本占比的期望为:



经过DYK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 DYK plus

在DYK算法中，我们设置了样本分类个数参数K，其思想就是将原本的特征空间划分为具有等量的K个样本的若干个子空间，我们将相邻子空间的任意样本之间进行K次线性插值，既K-path linear interpolation，以此增加有效的观测样本以及减小均匀噪音。这种方法对于参数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DYK方法难以达到很好的效果时，可以考虑换一种插值的思路，结合适合样本的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DYK plus。下面我们详细阐述下DYK 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：



其中，表示其类的中心距离空间最小点p最近的类，类的中心，d=1,…,k，为第d个类中样本个数，满足，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，对所有的，使其分别与中所有样本之间进行 次线性插值，插值样本个数为 ，并且插值的样本之间也是等距的。

|  |
| --- |
| **Algorithm2** DYK plus方法 |
| Require:单位距离填充参数  Require: 距离度量方法  Require: 特征空间最小点P |
| 对样本观测数据进行聚类，聚类个数为K，并根据所有类中心与p的距离得到  For d=1,…,k-1 do    确定的下一个进行插值的类  For i = 1,…, do  For j = 1,…, do  在间等距且线性插入个虚拟样本。  **end**  **end**  **end** |

## 6.2 the proof of DYK plus

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

与DYK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 中存在的样本量分别为，进行 次线性插值，每次插入的样本数量为m，插值后的均匀噪音为：



易得，，故插值后的均匀噪音减小。从极限的角度来看，DYK plus方法依旧可以使得均匀噪音减少至。

噪音误差大于0.5的样本占比的期望为:



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

从计算结果来看，DYK plus方法能达到的优化效果与DYK方法一致，但是涉及更多的计算。

**Remark：**DYK 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。DYK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DYK方法上提出了另一种基于类与类之间的线性插值方法——DYK plus，该方法也可以达到很好的优化效果。总的来说，我们发现DYK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.

Taylor L, Nitschke G. Improving deep learning with generic data augmentation[C]//2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2018: 1542-1547.

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.

Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

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[J]. 2005.

Cubuk E D, Zoph B, Mane D, et al. Autoaugment: Learning augmentation policies from data[J]. arXiv preprint arXiv:1805.09501, 2018.

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.