计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500145

不平衡油耗数据的区间预测方法

陈静杰[1,2,3]　王　琨[2,4]
1 中国民航大学电子信息与自动化学院　天津 300300
2 中国民航环境与可持续发展中心(智库)　天津 300300
3 综合交通大数据应用技术国家工程实验室　天津 300300
4 中国民航大学计算机科学与技术学院　天津 300300

摘　要　对油耗数据进行区间预测时,数据的不平衡性会导致一般的区间预测方法得到的预测区间质量较低。针对上述问题,提出了基于SMOTE-XGBoost算法的区间预测模型。采用SMOTE算法增加训练集中少数类样本的数量,消除了训练集数据的不平衡性;对XGBoost算法的分位数损失函数进行改进,平滑其一阶导数原点周围的小区域,解决了分位数损失函数对树分裂的影响;通过训练区间预测模型,得到预测区间的上下界。最后基于QAR数据集进行对比实验,结果表明,该方法使预测区间具有较高的区间覆盖率和较窄的区间宽度,提高了预测区间的质量。

关键词:不平衡数据;区间预测;SMOTE;XGBoost;油耗;Quick Access Recorder(QAR)数据
中图法分类号　TP391

Interval Prediction Method for Imbalanced Fuel Consumption Data

CHEN Jing-jie[1,2,3] and WANG Kun[2,4]
1 College of Electronic Information and Automation,Civil Aviation University of China,Tianjin 300300,China
2 Research Center for Environment and Sustainable Development of CAAC,Tianjin 300300,China
3 National Engineering Laboratory for Integrated Traffic Data Application Technology,Tianjin 300300,China
4 College of Computer Science and Technology,Civil Aviation University of China,Tianjin 300300,China

Abstract　Fuel consumption data is imbalanced,which leads to the lower quality prediction interval. Aiming at this problem,an interval prediction model based on SMOTE-XGBoost algorithm is proposed. From the perspective of oversampling,the SMOTE algorithm is used to increase the number of minority samples in the training set,so that the imbalance of data in the training set is eliminated. For the interval prediction task,the quantile loss function is used as the loss function of the XGBoost algorithm. At the same time,by smoothing the small area around the origin of its first derivative,the quantile loss function is improved to solve the problem that the quantile loss function causes the tree in the XGBoost algorithm to not split. Based on the above work,the XGBoost algorithm and SMOTE algorithm are combined to train the interval prediction model,and finally the upper and lower bound of the prediction interval are obtained respectively. Conducting experiments based on the QAR data set,the experiment results indicate that compared with other methods,this method makes the prediction interval have higher interval coverage and narrower interval width,which improves the quality of the prediction interval.

Keywords　Imbalanced data,Interval prediction,SMOTE,XGBoost,Fuel consumption,Quick Access Recorder(QAR) data

1　引言

为了减少国际民航每年CO_2的排放总量,国际民航组织(ICAO)提出了国际航空碳抵消和削减计划(CORSIA)[1]。要求飞机运营商聘请第三方对其年度排放报告进行核查,即检查其报告的油耗数据的合理性。但是核查单位作为独立的第三方并没有企业全部的数据,因此需要建立一个基于飞机运营商的实际历史油耗数据的油耗预测模型。在建立油耗预测模型时,目前国内外通常采用的方法是点预测,其预测的结果为航程对应的油耗的点值[2],但由于在一定区间内,大于预测点值的数据也是合理的,所以采用区间预测给出合理的预测范围是更为可靠的方法。

传统的区间预测方法如最小二乘法(OLS)回归法[3]、分位数回归法(quantile regression)[4]等可能会受到数据分布假

到稿日期:2020-05-28　返修日期:2020-10-27　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:中美绿色航线项目(GH201661279)
This work was supported by the Sino-US Green Route Pilot Program(GH201661279).
通信作者:陈静杰(jjchen@cauc.edu.cn)

设的限制，导致其预测区间质量不高[5]。目前能得到较高质量的预测区间的是集成学习方法，具体有随机森林(Random Forest，RF)[6]、梯度提升决策树(Gradient Boosting Decision Tree，GBDT)[7]等。RF是基于Bagging思想的集成学习方法，由多个没有关联的决策树组成，它在分类问题上表现良好[8]，常被用于处理分类问题。同时RF在区间预测方面能得到较高质量的区间[9]。GBDT在RF的基础上有进一步提升，它的基学习器为分类回归树，预测的准确度较高[10]，但基分类回归树之间存在依赖，难以并行训练数据。XGBoost(eXtreme Gradient Boosting)是由Chen[11]提出的基于GBDT的优化集成算法，它在特征粒度上实现并行计算，进行缓存优化，具有更高的训练效率，同时XGBoost增加了二阶梯度，使得预测区间具有更高的质量。由于分位数损失函数会导致XGBoost算法的树不分裂，使得模型表现不佳，故该算法没有被广泛应用于区间预测任务。

此外，油耗数据集中某些航程的油耗样本数量远远小于其他航程，因此该数据集通常是不平衡性数据集。一般情况下，当数据集中各个类别的样本数量大致相等时，该数据集为平衡数据集；当样本数量比例相差很大时，该数据集为不平衡数据集[12]。不平衡数据集中样本数较少的类称为少数类，样本数较多的称为多数类[13]。由于少数类样本提供的信息过少，因此算法会偏向于大多数群体，导致预测结果并不精确[14]。如果单纯地通过增加总体样本来增加少数类样本的数量，会消耗大量样本，增加运算成本。因此，提高少数类样本的分类精度是解决数据不平衡问题的关键。与随机过采样方法使用重复的真实数据进行过采样不同，SMOTE(Synthetic Minority Oversampling Technique)方法[15]基于插值合成新样本，避免了随机实例的重复，降低了过拟合的可能性。因此SMOTE算法常被用于处理数据不平衡的问题。

针对上述问题，本文提出了基于SMOTE-XGBoost算法的区间预测模型。其通过SMOTE算法对油耗数据进行处理，增加了训练集中少数类样本的数量，减少了不平衡数据对区间预测的影响；为实现区间预测，采用分位数损失作为XGBoost算法的损失函数，同时平滑其一阶导数原点周围的小区域，解决了分位数损失函数导致XGBoost算法中的树不分裂的问题。最后训练区间预测模型，构造预测区间。通过对比实验可知，本文提出的方法比现有的算法获得了更高质量的区间，增强了对油耗数据进行合理性检查的可靠性。

2 不平衡数据区间预测模型的构建

不平衡数据区间预测模型首先对QAR数据作预处理，得到用于构造预测区间的航程油耗数据，从过采样的角度通过SMOTE算法平衡训练集，然后使用XGBoost算法作为主要的机器学习的方法来构建区间。XGBoost采用改进的分位数损失函数，通过网格交叉验证(GridSearchCV)的方式得到最优的参数组合。本文随机选取数据，以8∶2的比例得到训练集和测试集，并对训练集进行交叉验证和参数调试，以完成模型的建立。最后利用测试集对模型的预测效果进行检测。

2.1 SMOTE算法

SMOTE[15]是Chawla等提出的应用于处理不平衡数据的过采样技术。油耗数据集以航程为分类依据。SMOTE算法首先确定油耗数据集中所有少数类样本的k近邻，然后在少数类样本与其k个最近邻之间的直线上插入人工少数样本，将合成的新样本添加到数据集中，直到数据集达到平衡。SMOTE算法合成新样本过程如图1所示。

算法流程：

(1)对于油耗数据集的少数类中的每一个样本x_i，$i\in\{1,\cdots,T\}$，计算它们到其他少数类样本的欧氏距离，确定其k近邻，记为x_{inn}，$nn\in\{1,\cdots,k\}$(k默认为5[16])。

(2)从x_i的k近邻x_{inn}中随机选择一个样本作为合成样本的辅助样本，通过式(1)得到合成样本x_{inew}的属性值，即新样本的油耗值$x_{inew,att}$：

$$x_{inew,att}=x_{i,att}+\gamma(x_{inn,att}-x_{i,att}) \tag{1}$$

其中，$x_{i,att}$和$x_{inn,att}$分别为x_i和x_{inn}的油耗值，γ是[0,1]之间的随机数。

(3)将上一步重复n次，得到n个合成样本(n为过采样量)。

(4)对每一个x_i重复步骤(2)－(3)，最终得到$n\times T$个新样本。

(5)将新样本添加到数据集中。

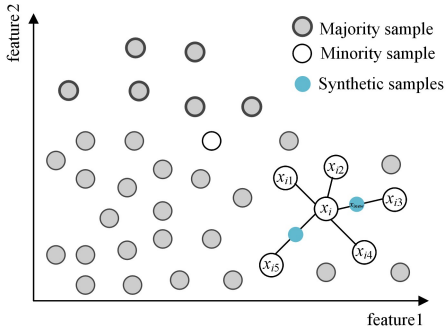


图1 SMOTE算法图示

Fig. 1 SMOTE algorithm diagram

2.2 XGBoost算法

XGBoost是GBDT算法的高效实现[11]，与GBDT相比，XGBoost一方面使用二阶泰勒展开来逼近损失函数，快速优化目标，另一方面在目标函数中加入正则化项，提高了模型的泛化能力。XGBoost回归树集成模型如图2所示。

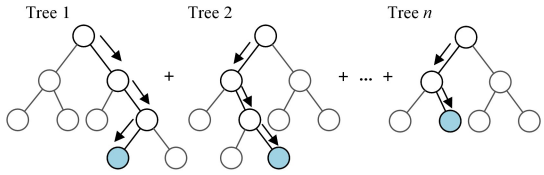


图2 XGBoost回归树集成模型

Fig. 2 XGBoost regression tree integration model

2.2.1 XGBoost目标函数优化

XGBoost的目标函数由损失函数和正则项组成：

$$obj^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t)})+\sum_{k=1}^{K}\Omega(f_k) \tag{2}$$

其中，y_i是真实的油耗值，$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t(x_i)$是第$t$轮预测结果。$\hat{y}_i^{(t-1)}$是第$t-1$轮生成的决策树预测结果，$f_t(x_i)$是第$t$次迭代产生的决策树函数。

对损失函数进行二阶泰勒展开得：

$$obj^{(t)}=\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{(t-1)})+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)+c \tag{3}$$

其中，$g_i=\alpha_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$，$h_i=\alpha^2_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$分别为损失函数的一阶和二阶导数，$c$ 为常数。

由于$\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)})+c$对目标函数求最优解没有影响，故目标函数简化为：

$$obj^{(t)}\approx\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right] \tag{4}$$

用 I_j 来表示第 j 个叶子里样本集合：

$$I_j=\{i|q(x_i)=j\} \tag{5}$$

通过叶子中的一个分数向量定义树，并通过一个叶子索引映射函数将样本映射到叶子：

$$f_t(x)=\omega_{q(x)},\omega\in\boldsymbol{R}^T,q:\boldsymbol{R}^d\rightarrow\{1,2,\cdots,T\} \tag{6}$$

其中，ω 为树叶的权重序列，q 为树的结构，$q(x)$表示样本 x 所在树叶的位置。

用基分类回归树叶子节点的个数 T 和树叶权重 ω_j 的平滑程度来描述模型的复杂度：

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{7}$$

其中，γ 为新叶子节点的复杂度代价，$\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$表示叶子权重的平滑程度。

将目标函数按每个叶子重新组合：

$$\begin{aligned}obj^{(t)}&=\sum_{j=1}^{T}\left[(\sum_{i\in I_j}g_i)\omega_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)\omega_j^2\right]+\gamma T\\&=\sum_{j=1}^{T}\left[G_j\omega_j+\frac{1}{2}(H_j+\lambda)\omega_j^2\right]+\gamma T\end{aligned} \tag{8}$$

其中，$G_j=\sum_{i\in I_j}g_i$，$H_j=\sum_{i\in I_j}h_i$。

对于固定树结构，叶子 j 的最优权重 ω_j^* 由顶点公式可得：

$$\omega_j^*=-\frac{G_j}{H_j+\lambda} \tag{9}$$

将式(9)代入式(8)，得到最优目标函数：

$$obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{10}$$

式(10)可用作衡量树结构 q 质量的评分函数。该分数越小，即目标函数值越小，代表树的结构越好。

树学习的一个关键问题是找到最优的树结构。通常不可能枚举所有可能的树结构 q，取而代之的是一种贪心法，它需要遍历所有特征中可能的分裂点位置。对一个叶子节点进行分裂，分裂后的增益为：

$$Gain=\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}-\gamma \tag{11}$$

其中，第一项为左子树的分数，第二项为右子树的分数，第三项为分裂的分数。$Gain$ 值越大，分裂后损失函数减小越多。因此计算所有候选特征后选取 $Gain$ 值最大的特征进行分裂，同时如果增益为负值，则停止分裂。

2.2.2 XGBoost 分位数损失函数的改进

XGBoost 是一种框架算法，针对不同的任务使用不同的损失函数，同时它支持自定义损失函数，根据目标函数优化过程，损失函数满足二阶可导即可。在使用时需提供损失函数的一阶和二阶导数。

机器学习中分位数损失常用于区间预测。利用分位数损失无须对数据进行先验处理，即可预测某一分位数的水平值。给定预测值$\hat{y}_i$和真实油耗值 y_i，对于分位数 q 的分位数损失函数的定义如下[17]：

$$L_q(y,\hat{y})=\sum_{i=y_i<\hat{y}_i}(1-q)|y_i-\hat{y}_i|+\sum_{i=y_i\geqslant\hat{y}_i}(q)|y_i-\hat{y}_i| \tag{12}$$

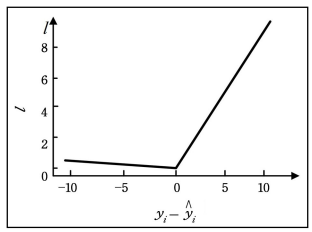
分位数损失的一阶导数为：

$$g_i=\alpha_{\hat{y}_i}l(y_i,\hat{y}_i)=\begin{cases}1-q, & y_i-\hat{y}_i<0\\-q, & y_i-\hat{y}_i\geqslant 0\end{cases} \tag{13}$$

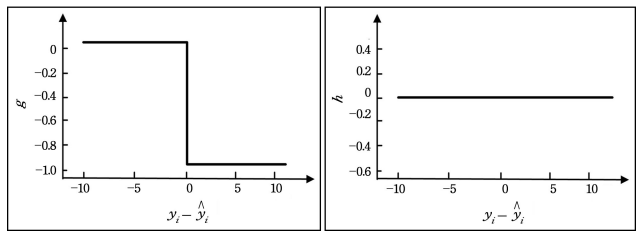
二阶导数为：

$$h_i=0 \tag{14}$$

分位数损失函数及其一阶和二阶导数如图 3 所示。



(a)损失函数 $q=0.95$



(b)一阶导数 $g=a_{\hat{y}_i}l$　　(c)二阶导数 $h=a^2_{\hat{y}_i}l$

图 3 分位数损失及其一阶导数和二阶导数

Fig. 3 Quantile loss function and its first and second derivative

从图 3 可看出分位数损失函数的二阶导数 h 恒为 0，使得式(11)中的 $Gain$ 值为负，则树不再分裂，导致模型进行区间预测的效果不佳。根据数据集中样本数值的大小，得到预测区间的上下界为两条油耗值不同的直线。

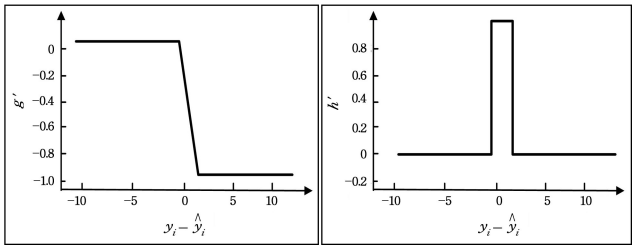
因此，对接近分位数的点，通过平滑原点周围的小区域 Δ，对分位数损失函数进行改进，此时二阶导数便不全为 0。

$$g_i'=\alpha_{\hat{y}_i}l(y_i,\hat{y}_i)=\begin{cases}1-q, & y_i-\hat{y}_i<-(1-q)\Delta\\-x/\Delta, & -(1-q)\Delta\leqslant y_i-\hat{y}_i<q\Delta\\-q, & y_i-\hat{y}_i\geqslant q\Delta\end{cases} \tag{15}$$

$$h_i'=\begin{cases}\dfrac{1}{\Delta}, & -(1-q)\Delta\leqslant y_i-\hat{y}_i<q\Delta\\0, & \text{其他}\end{cases} \tag{16}$$

由此，分位数损失函数的一阶及二阶导数如图 4 所示。

当二阶导数不全为 0 时，$Gain$ 值不恒为负，XGBoost 算法的树可分裂，因此可以得到预期的预测区间。



(a)一阶导数 $g'=a_{\hat{y}_i}\ l$　　(b)二阶导数 $h'=a^2_{\hat{y}_i}\ l$

图 4　改进分位数损失的一阶导数和二阶导数

Fig. 4　Improved first and second derivative of quantile loss function

2.3　模型的结构

不平衡油耗数据的区间预测模型结构如图 5 所示。该模型主要分为数据预处理、SMOTE 算法平衡数据集、XGBoost 算法构造预测区间以及区间预测结果分析 4 个部分。

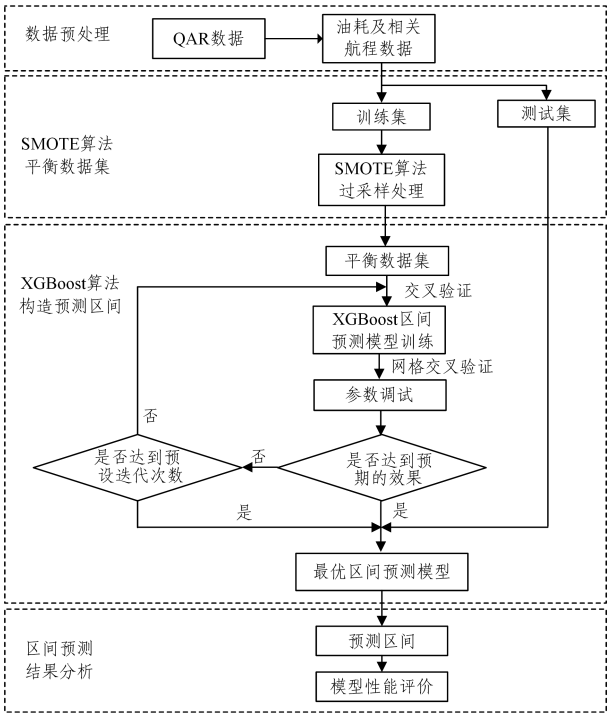


图 5　不平衡油耗数据的区间预测模型

Fig. 5　Interval prediction model of imbalanced fuel consumption data

首先从 QAR 数据集中获取需要的油耗及相关航程的数据。然后通过 SMOTE 算法合成新的样本，增加油耗数据集中少数类样本的数量，使训练集中少数类样本与多数类样本数量相同，即数据集达到平衡。

XGBoost 算法采用改进的分位数回归函数来构造区间，其学习过程是一个迭代过程，每次迭代对应于叶节点的分裂。每次分裂相当于将属于一个叶节点的训练样本分配给分裂的两个新叶节点。为了建立不平衡数据的区间预测模型，需要对模型中的各种参数进行调整，使模型参数达到最优。

通过对训练集进行模型训练和参数调试，完成区间预测模型的建立。最后，利用测试集得到预测区间，并采用多个评价指标对模型进行评价。

3　实验分析

3.1　实验数据

本实验使用 A330 机型 2013 年的部分 QAR 数据，从中取得相关航程和油耗数据共 1 360 组。不同航程对应的油耗样本数如图 6 所示。可以看出，航程为 1 178 公里的油耗样本数最多，有 305 个，航程为 1 133 公里对应的样本数最少，为 8 个。这种样本数据量的差距造成了数据的不平衡，其原因是航班数的差别，航程 1 178 公里对应的航班为北京到上海，一段时间内航班较多，而 1 133 公里对应的航班为北京到武汉，同样的时间内航班数较少。

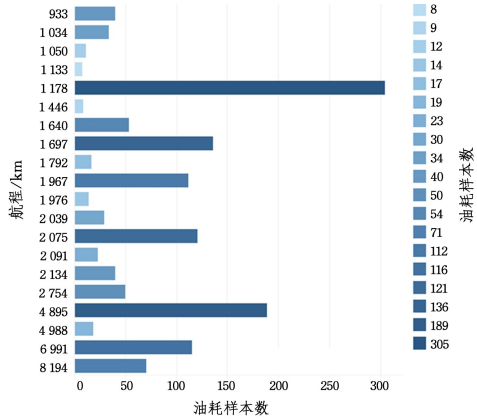


图 6　不同航程对应油耗样本数的统计图

Fig. 6　Statistical chart of fuel consumption samples corresponding to different voyage

本实验以 8∶2 的比例得到训练集和测试集，训练集样本总数为 1 088，测试集样本总数为 272。

3.2　SMOTE 处理不平衡油耗数据的结果

在油耗数据集中，不同航程对应油耗样本数量均不相同。实验中将训练集中所有航程对应油耗样本的数量都变为与样本数最多的类相同。同时测试集保持原来的油耗样本不变。最终得到训练数据集样本总数为 4 880，测试集样本总数仍为 272。

3.3　实验数据参数设置

为了建立不平衡数据的区间预测模型，需要对模型中的参数进行调整，使模型效果达到最优。

调整模型参数的步骤如下：1)在初始化参数值的基础上调整学习率(*learning_rate*)和决策树的数目(*n_estimators*)。固定学习率为 0.2，使用 GridSearch 工具遍历可能的决策树数目的候选值并找到最佳值。2)调整树最大深度(*max_depth*)和叶子节点最小权重(*min_child_weight*)参数。同样使用 GridSearch 工具遍历可能的候选参数并找到最佳参数。3)调整分裂收益阈值(*gamma*)参数，可以先选择较大粒度值来确定最佳参数范围，然后选择较小的值进行微调。4)使用相同的方法对其他参数 *subsample*，*colsample_bytree* 以及 Δ 进行调试。

本实验设置 *learning_rate* 为 0.2；*n_estimators* 为 570；*max_depth* 为 3；*min_child_weight* 为 0；*gamma* 为 0；*subsample* 为 0.6；*colsample_bytree* 为 0.6；Δ 区间下界设为 81，区间上界设为 100。

3.4 模型评价指标

预测区间的质量通常由两种不同的度量来评估，即预测区间覆盖概率和预测区间宽度[18]。预测区间覆盖概率(Prediction Interval Coverage Probability，PICP)是指所构造的预测区间捕获实际目标变量的能力，即落在区间内的油耗样本数占总样本数的百分比。$PICP$ 表示如下：

$$PICP=\frac{1}{N}\sum_{i=1}^{N}C_i, C_i=\begin{cases}1, & y_i\in[L_i,U_i]\\ 0, & y_i\notin[L_i,U_i]\end{cases} \tag{17}$$

其中，n 是测试集中的油耗样本数，y_i 是真实油耗值，U_i 和 L_i 分别是第 i 个预测区间的上下限。

通过构造宽的预测区间可以很容易地获得很高的预测区间覆盖率，但是这样的预测区间没有传递关于潜在目标变量变化的信息，且在实际中没有用处。因此，还必须通过预测区间归一化平均宽度(Prediction Interval Normalized Average Width，PINAW)来评估其质量，$PINAW$ 的公式如下：

$$PINAW=\frac{1}{rN}\sum_{i=1}^{N}(U_i-L_i) \tag{18}$$

其中，r 表示目标在整个预测区间获得的最大值和最小值的范围，即最大观测值与最小观测值的差。

在实际应用中，人们希望在 $PICP$ 达到目标区间置信度要求的同时，能够获得较窄的 $PINAW$。为了综合评估两个度量，一个由 $PICP$ 和 $PINAW$ 组成的综合指标覆盖宽度准则(Coverage Width Criterion，CWC)被提出。

$$CWC=PINAW*(1+\nu(PICP)e^{(-\eta(PICP-\mu))}) \tag{19}$$

其中，$\nu(PICP)=\begin{cases}0, & PICP\geqslant\mu\\ 1, & PICP<\mu\end{cases}$，$\mu$ 和 η 分别为控制 CWC 大小的超参数，η 值设置在 10 到 100 之间，以惩罚无效的预测区间，本文中 η 设置为 50，μ 为其构造区间的标称置信水平。

3.5 区间预测结果的比较

当目标置信度为 90%时，SMOTE-XGBoost 算法的区间预测模型的预测效果如图 7 所示。图 7 显示了从 933 公里到 8194 公里的航程对应的油耗预测区间，其中蓝色直线为运用 XGBoost 算法求得的油耗的预测值，点为油耗观察值，淡蓝色区域为油耗数据 90%的预测区间。由于油耗的值跨度较大，较短航程的预测效果不明显，故局部放大图如图 8 所示。

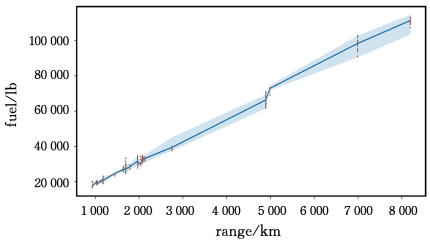


图 7 区间预测效果图(电子版为彩色)

Fig. 7 Interval prediction results

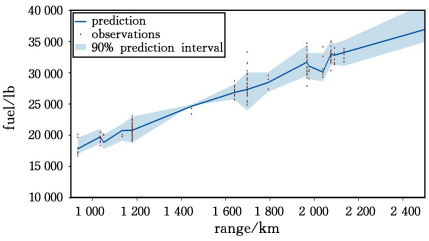


图 8 区间预测效果局部放大图

Fig. 8 Local magnification of interval prediction results

从图 8 可以看出，区间预测结果上下界高度对称，XGBoost 算法的预测值接近上下限的平均值。

为了进一步验证不平衡油耗预测区间的质量，选取常用的区间预测方法进行对比实验。对比方法分别为分位数回归、RF 和 GBDT。在此基础上，分别对不平衡的训练集以及通过 SMOTE 算法平衡的训练集进行学习，构建预测区间。本实验中目标置信度设为 0.9，采用 $PICP$，$PINAW$ 和 CWC 作为评价标准。在区间覆盖率达到目标置信度的基础上，预测区间的归一化平均宽度越小越好，综合指标覆盖宽度准则作为综合上述两个指标的准则，其值越小，则区间质量越高。不平衡训练集下各方法的预测区间质量如表 1 所列，平衡训练集下各方法的预测区间质量如表 2 所列。

表 1 不平衡训练数据的预测区间评价

Table 1 Evaluation of prediction interval of imbalanced training data

Algorithm	$PICP$	$PINAW$	CWC
Quantile regression	0.821	0.145	7.516
RF	0.886	0.041	0.123
GBDT	0.900	0.058	0.058
XGBoost	0.903	0.056	0.056

表 2 平衡训练数据的预测区间评价

Table 2 Evaluation of prediction interval of balanced training data

Algorithm	$PICP$	$PINAW$	CWC
Quantile regression	0.921	0.155	0.155
RF	0.918	0.122	0.122
GBDT	0.900	0.053	0.053
XGBoost	0.900	0.052	0.052

从表 1 可以看出，分位数回归、RF 在使用不平衡数据集进行训练时，$PICP$ 均未达到目标置信度 0.9，这使得 CWC 分别为值较大的 7.516 和 0.123，预测区间质量较低。而 GBDT 以及 XGBoost 在使用不平衡训练集时已经可以取得较好的效果，$PICP$ 均达到 0.9 以上，同时 $PINAW$ 分别为较小的 0.058 和 0.056，CWC 值分别与各自的 $PINAW$ 值相同。这是根据式(19)，当 $PICP$ 大于或等于构造区间的标称置信水平 μ 时，$\nu(PICP)=0$，导致 CWC 与 $PINAW$ 值相同。

在表 2 中使用 SMOTE 平衡数据集进行训练，各方法都有较好的效果，4 种算法均达到了目标置信度 0.9，并且 $PINAW$ 较小，使得 CWC 值也较小且均与 $PINAW$ 相等。

比较表 1 和表 2 可以看出，4 种算法使用平衡训练集进行训练得到的 CWC 值都有所减小，其结果均比不平衡训练集的结果好。同时基于 SMOTE-XGBoost 算法构建的区间，在 $PICP$ 达到目标置信度的基础上，得到最小的 CWC 值，模型表现是最好的。根据上述比较，结合 SMOTE 算法和 XGBoost 算法的区间预测模型能得到高质量的预测区间。

结束语 本文针对不平衡数据集对预测区间质量的影响，提出了基于 SMOTE-XGBoost 算法的区间预测模型。其利用 SMOTE 算法增加了油耗训练集中少数类样本的数量，使训练集多数类与少数类的样本量达到平衡。同时，针对 XGBoost 算法树分裂的特性改进分位数损失函数，平滑其一阶导数原点周围的小区域。将 XGBoost 算法与 SMOTE 算法相结合来训练区间预测模型，最终得到预测区间的上界。

本文采用 PICP，PINAW，CWC 为模型评价指标。实验结果证明，与其他区间预测方法相比，当目标置信度为 0.9 时，本文提出的方法能够获得 0.900 的区间覆盖率，0.052 的预测区间归一化平均宽度，以及 0.052 的综合指标覆盖宽度准则值，可得到高质量的区间，增强了对油耗数据进行合理性检查的可靠性。

参 考 文 献

[1] MICHAELOWA A. Tackling CO_2 emissions from international aviation: challenges and opportunities generated by the market mechanism 'CORSIA'[J]. EDA Insight, 2016, 2(11): 1-7.

[2] STROUHAL M. CORSIA-Carbon Offsetting and Reduction Scheme for International Aviation[J]. MAD-Magazine of Aviation Development, 2020, 8(1): 21-26.

[3] VILAR J, ANEIROS G, RAÑA P. Prediction intervals for electricity demand and price using functional data[J]. International Journal of Electrical Power & Energy Systems, 2018, 96(3): 457-472.

[4] NOWOTARSKI J, WERON R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging[J]. Computational Statistics, 2015, 30(3): 791-803.

[5] MENG Y, ZHANG B, YAN Y M. Prediction Interval Estimation Model of User Concurrent Requests for Cloud Service in Cloud Environment[J]. Chinese Journal of Computers, 2017, 40(2): 378-396.

[6] ROY M H, LAROCQUE D. Prediction intervals with random forests[J]. Statistical Methods in Medical Research, 2020, 29(1): 205-229.

[7] VERBOIS H, RUSYDI A, THIERY A. Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting [J]. Solar Energy, 2018, 173: 313-327.

[8] PENG Z, WANG L Q, GUO H. Parallel Text Categorization of Random Forest[J]. Computer Science, 2018, 45(12): 148-152.

[9] ZHANG H, ZIMMERMAN J, NETTLETON D, et al. Random forest prediction intervals[J]. The American Statistician, 2020, 74(4): 392-406.

[10] HUANG J, ZHU L, FAN B, et al. Large-Scale Price Interval Prediction at OTA Sites[J]. IEEE Access, 2018, 6: 69807-69817.

[11] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[12] KAUR H, PANNU H S, MALHI A K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions[J]. ACM Computing Surveys (CSUR), 2019, 52(4): 1-36.

[13] GUO H X, LI Y J, SHANG J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems With Applications, 2016, 73: 220-239.

[14] ZHENG Z, CAI Y, LI Y. Oversampling method for imbalanced classification[J]. Computing and Informatics, 2016, 34(5): 1017-1037.

[15] 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.

[16] 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.

[17] KOENKER R, BASSETT J G. Regression quantiles. Econometrica[J]. Journal of the Econometric Society, 1978, 46(1) 1: 33-50.

[18] QUAN H, KHOSRAVI A, YANG D, et al. A survey of computational intelligence techniques for wind power uncertainty quantification in smart grids[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(11): 4582-4599.

CHEN Jing-jie, born in 1967, Ph.D, professor. His main research interests include energy efficiency management and carbon emission control in civil aviation transportation.