

江煦宁, 薛雅慧, 王一菲. 基于 SMOTE 改进方法的交通事故风险预测模型[J]. 智能计算机与应用, 2025, 15(7): 42-47.
DOI: 10.20169/j.issn.2095-2163.25021902

基于 SMOTE 改进方法的交通事故风险预测模型

江煦宁¹, 薛雅慧², 王一菲³

(1 同济大学 数学科学学院, 上海 200092; 2 同济大学 物理科学与工程学院, 上海 200092;

3 同济大学 国豪书院, 上海 200092)

摘要: 近年来, 交通事故发生率居高不下, 给社会稳定和经济发展造成了严重影响。交通事故预测属于不平衡数据的二分类问题, 具有数据量大、样本不平衡率高、计算强度大等问题。采集 G15 上海路段高速公路交通数据, 采用合成少数类样本过采样技术 (Synthetic Minority Oversampling Technique, SMOTE) 改进方法和机器学习、深度学习模型结合, 改善 SMOTE 方法合成数据时的盲目性和不同分类模型处理不平衡数据时产生的过拟合问题, 分类探究了不同模型对交通事故风险预测的适用性和鲁棒性, 并针对实验数据得出结论。实验表明, 属性分类合成过采样 (Attribute-Synthetic Minority Oversampling Technique, ASMOTE) 和极致梯度提升 (eXtreme Gradient Boosting, XGBoost), 改进合成过采样 (Improved Synthetic Minority Oversampling Technique, ISMOTE) 和随机森林 (Random Forest, RF) 两种结合模型在交通事故预测领域具有一定优势。运用 SMOTE 改进方法有效避免了交通事故样本误分类率高的问题, 提升了交通事故风险预测的准确性, 为实现交通事故预测提供解决策略与方向。

关键词: 交通事故; 风险预测; 不平衡数据; 过采样; SMOTE 改进方法

中图分类号: TP399

文献标志码: A

文章编号: 2095-2163(2025)07-0042-06

Prediction model of traffic accident risk based on improved SMOTE methods

JIANG Xuning¹, XUE Yahui², WANG Yifei³

(1 School of Mathematical Sciences, Tongji University, Shanghai 200092, China;

2 School of Physics Science and Engineering, Shanghai 200092, China;

3 Guohao College, Tongji University, Shanghai 200092, China)

Abstract: In recent years, the incidence of traffic accidents has remained high, causing huge losses to social stability and economic development. Traffic accident prediction belongs to the binary classification problem of imbalanced data, featuring large amounts of data, a high sample imbalance ratio, and high computational intensity. Traffic data on the G15 Shanghai section of the expressway was collected. The improved methods of the Synthetic Minority Oversampling Technique (SMOTE) were combined with machine learning and deep learning models to address the blindness of the SMOTE algorithm in synthesizing data and the overfitting problem that occurs when different classification models handle imbalanced data. The applicability and robustness of different models for traffic accident risk prediction were explored through classification, and conclusions were drawn based on the experimental data. The experiments show that two combined models, namely the Attribute - Synthetic Minority Oversampling Technique (ASMOTE) and eXtreme Gradient Boosting (XGBoost), as well as the Improved Synthetic Minority Oversampling Technique (ISMOTE) and Random Forest (RF), have certain advantages in the field of traffic accident prediction. By applying improved SMOTE algorithms, these models effectively avoid the problem of a high misclassification rate of traffic accident samples, improve the accuracy of traffic accident risk prediction, and provide solutions and directions for achieving traffic accident prediction.

Key words: traffic accident; risk prediction; unbalanced data; Oversampling; improved SMOTE algorithms

0 引言

交通事故是对生命安全的重大威胁之一。根据

国家统计局数据, 近年交通事故发生率居高不下, 截至 2022 年, 交通事故高达 256 409 起, 死亡人数约 6 万人, 直接财产损失达 123 926 万元, 不仅致使人员

基金项目: 2024 年上海市级创新训练项目 (S202410247244)。

作者简介: 江煦宁 (2004—), 女, 本科生, 主要研究方向: 数学与应用数学。Email: yangxushanli@tongji.edu.cn; 薛雅慧 (2004—), 女, 本科生, 主要研究方向: 应用物理学; 王一菲 (2004—), 女, 本科生, 主要研究方向: 数学与应用数学 (强基计划)。

收稿日期: 2025-02-19

哈尔滨工业大学主办 ◆ 学术研究与应用

伤亡和财产损失,还在一定程度上阻碍了经济发展。如果能根据环境因素对事故发生概率做出预测和判断,便可通过交通部署和应急救援资源配置等方式降低事故的伤亡和损耗。为此,要寻求高效的模型预测碰撞发生的次数或频率^[1],探究碰撞发生的影响因素,为减少交通事故的发生提供对策。

然而,交通事故因发生频率低导致数据具有高度不平衡特征。学习的碰撞数据样本量不足导致现有的建模方法如逻辑回归、随机森林和支持向量机(Support Vector Machine, SVM)等模型的预测性能显著下降^[2]。因此,基于大规模不平衡数据实现交通事故预测成为研究课题的关键。

传统的不平衡数据改进方法通过欠采样或过采样调整类分布^[3]。本文创新性探讨 SMOTE 改进方法与神经网络等模型的融合机制,基于 G15 高速公路上海路段的交通数据验证模型的有效性,并提出针对性解决策略,为实现交通事故预测提供方向。

1 文献综述

随着道路检测系统的完善,实时碰撞风险预测方法也在不断发生变化^[4-5]。通常预测方法可以分为统计方法和机器学习方法。统计方法通过对数据集的整理与分析,包括对数线性模型、贝叶斯逻辑模型等,得到影响交通事故发生概率的主要因素^[6-7]。但由于交通事故数据的不平衡特征,统计方法在事故预测上表现不佳。为了实现更高的预测精度, SVM、XGBoost 等机器学习方法广泛用于交通预测模型中^[8-9]。Pande 等^[10]将神经网络模型用于高速公路事故风险预测,其不同路段的效果不同^[10]; Yu 和 Abdel - At^[11]使用分类与回归树模型(Classification and Regression Tree, CART)选择变量,引入 SVM 模型评估实时碰撞风险,并与贝叶斯逻辑回归模型进行比较。近年来,深度学习方法与集成学习方法在不平衡数据处理方面表现出优异的性能,一些学者也将其融入交通事故数据集的分析与处理中,例如 Schlögl^[12]采用平衡 Bagging 方法训练弱分类器,用于预测事故发生风险。

为了降低交通数据集的不平衡特征对建模精度的影响,避免丢失数据信息, SMOTE 过采样方法广泛应用于平衡数据。Basso 等^[13]使用 SMOTE 方法对高速公路路段的碰撞数据进行处理,发现基于 SMOTE 过采样的平衡数据集的模型预测性能较为优异。不少实时碰撞风险预测采用 SMOTE 方法进

行过采样,用于比较不同方法在事故风险预测方面的优劣。近年来, SMOTE 过采样各类改进方法与不同算法结合,广泛应用于多个领域的分类和预测中^[14-15]。SMOTEBagging 将 SMOTE 与 Bagging 集成分类器结合,提高分类器的分类性能和少数类样本预测的准确率^[16]。Kosolwattana 等^[17]提出了自检式自适应 SMOTE (A self-inspected adaptive SMOTE algorithm, SASMOTE)方法,通过自检消除生成样本不确定性的方法,生成高质量合成样本,在风险基因检测和先天性心脏病预测等医疗案例中实现了较为理想的预测性能; Ileberi 等^[18]将 SMOTE 与自适应增强技术(Adaptive Boosting, AdaBoost)结合,建立了基于机器学习的信用卡欺诈检测框架; 琚春华等根据信用卡交易数据的序列关系,结合 k 最近邻(k - Nearest Neighbor, kNN)、SMOTE 方法与长短期记忆网络(Long Short Term Memory networks, LSTM)构建信用卡欺诈检测模型^[19],对改善信用卡诈骗乱象提供了帮助;朱莉等引入中心偏移权重对 SMOTE 方法进行改进,结合变异麻雀搜索算法优化支持向量机(Variation Sparrow Search Algorithm - Support Vector Machine, VSSA-SVM)模型,建立变压器故障诊断分类器^[20],有效解决了变压器故障检测困难的问题;Feng 等^[21]选取 AdaBoost 和卷积神经网络(Convolutional Neural Network, CNN)结合,对 SMOTE 过采样后的交通事故数据进行测试,处理不平衡大数据能力显著提高。

本文基于 G15 高速公路上海路段交通数据,比较不同混合采样处理方法对交通事故风险预测的准确率和稳健性,为事故风险预测提供研究方向。

2 数据集准备和处理

本文以中国上海 G15 高速公路为研究对象,以电动汽车事故发生频率为因变量,数据包括道路几何特征、气象数据、交通运行数据等可能对事故发生产生影响的变量和通过综合交通管理数据库获取的事故发生情况。

为避免变量之间存在强相关性,对收集的 28 个变量进行 Pearson 相关性分析,设置相关阈值为 0.6,判断任意两个变量间是否高度相关,并将强相关变量选择性删除,最终得到包含 18 个变量的建模数据集。

数据包含变量属性见表 1。Feng 等^[21]对数据作了详细的描述与解释。

表 1 数据变量说明
Table 1 Description of the collected variables

变量名称	变量类型	变量说明
道路几何变量		
主要路段类型	离散型	1:主要路段远离坡段 2:主要路段靠近坡段
水平曲线长度	连续型	整段曲线长度(km)
水平曲线比例	连续型	曲线和切线对齐的比例(%)
垂直剖面类型	离散型	1:上坡 2:下坡 3:凸面曲线 4:凹面曲线 5:斜坡和曲线的结合路段
纵向坡段长度	连续型	垂直坡段总长度(km)
垂直曲线比例	连续型	总路段中垂直曲线长度比例(%)
设置护栏的中央隔离带宽度	离散型	1:宽度为 2 m 2:宽度大于 2 m
路灯数量(同一方向)	离散型	1:两盏路灯 2:大于两盏路灯
气象相关变量		
最大风速	连续型	风速最大值(m/s)
极端风速	连续型	特定条件下出现的非常高的风速情况(m/s)
风力等级	离散型	0:无风状态,风速低于 0.3 m/s 1:软风,风速在 0.3~1.6 m/s 之间 2:轻风,风速 1.6~3.4 m/s 3:微风,风速提升至 3.4~5.5 m/s
降雨量	连续型	一小时累计降水量(mm)
能见度	连续型	水平能见度(m)
天气	离散型	0:晴天(观测期间或观测前 1 小时以内) 1:有雨、尘雾或冰雹天气(观测前 1 小时内) 2:尘暴、沙暴或下雪天气(观测期间) 3:起雾或冰雾天气(观测期间) 4:雨天(观测期间)
交通运行变量		
交通流量	连续型	电动车的车流量(辆/h)
平均车速	连续型	每小时电动车平均车速(km/h)
速度变化	连续型	电动车每小时内速度变化
交通事故		
事故发生类型	离散型	0:一小时内未发生交通事故 1:一小时内发生一次或两次交通事故

已收集的所有变量,起始时间为 2021 年 4 月 2 日至 4 月 30 日,所有数据共 259 200 组,其中 259 093 组(99.96%)数据为未发生车祸类型,101 组(0.04%)为发生车祸类型。排除未有车辆经过

路段,最终选用 256 962 组数据进行实验,包括 107 组车祸案例和 256 855 组非车祸案例,不平衡比在 1 至 2 400 之间。

3 算法原理与步骤

3.1 平衡数据

3.1.1 ASMOTE 方法

SMOTE 方法中 k 近邻所使用的距离公式是欧氏距离,欧氏距离公式仅能计算数值属性。但实际的不平衡数据集存在离散属性(或称名义属性)。因此,在 k 近邻计算中仅使用欧氏距离公式会导致新样本无代表性。

本文在原有的算法基础上,引入 ASMOTE 方法,即在 k 近邻部分增加对离散属性的处理。在分类任务中,为了衡量两个样本 u 和 v 之间的相似性,定义其总体距离 $dist(u, v)$ 为各维度变量距离的累加和,公式如下:

$$dist(u, v) = \sum_{i=1}^n d_i(u_i, v_i) \quad (1)$$

其中, $d_i(u_i, v_i)$ 代表第 i 个变量在样本 u 和 v 之间的距离,其计算方式依据变量类型的不同而有所区别。

若第 i 个变量为离散变量或名义类变量时,采用 VDM(Value Difference Metric) 距离计算方法,通过该变量数量占比之差计算距离,公式如下:

$$d_i(u_i, v_i) = \left(\frac{N_{uc}}{N_u} - \frac{N_{vc}}{N_v} \right)^2 \quad (2)$$

其中, u_i 和 v_i 分别表示样本 u 和 v 第 i 个变量的数值; N_u 和 N_v 分别表示数据集中第 i 个变量数值为 u_i 和 v_i 的样本数; N_{uc} 和 N_{vc} 分别表示该类别中第 i 个变量数值为 u_i 和 v_i 的样本数。

若第 i 个变量为连续变量或数值类变量时,采用欧式距离计算公式计算距离,公式如下:

$$d_i(u_i, v_i) = (u_i - v_i)^2 \quad (3)$$

根据距离属性选择不同的距离计算方法,有助于生成数据更接近真实数据特征,提升合成样本的有效性。

3.1.2 ISMOTE 方法

为了降低 SMOTE 方法生成样本时噪声样本的干扰,解决少数类样本内不均衡的缺陷,ISMOTE 方法在 SMOTE 方法基础上,新增中心偏移权重算法(Central Offset Weight, COW),进一步对样本筛选去噪。

1) 使用 Tomek links 欠采样技术去除少量少数类样本;

2) 使用 COW 去除少数类样本:

(1) 选取少数类样本 x_i , 计算 k 近邻的区域中心

$C_k(X_m)$, 公式如下:

$$C_k(X_m) = \frac{1}{k} \sum_{q \in N_k(X_m)} X_q \quad (4)$$

其中, $N_k(X_m)$ 为 X_m 的 k 近邻的集合, $X_q = (X_{q1}, X_{q2}, \dots, X_{qm}) \in R^m$ 。

(2) 随着参数 k 的自增,用 $\sigma(X_m)$ 来衡量 $C_k(X_m)$ 的迁移量,公式如下:

$$\sigma_i(X_m) = d(C_i(X_m), C_{i+1}(X_m)) \quad (5)$$

其中, $i = 1, 2, \dots, k-1$, $\sigma_i(X_m)$ 通常在稀疏区域比在密集区域有更大的值。

(3) 采用中心迁移量的绝对误差表示 k 近邻中心位置变化的程度,公式如下:

$$Cow = \sum_{i=1}^{k-2} |\sigma_{i+1}(X_m) - \sigma_i(X_m)| \quad (6)$$

(4) 设定阈值,选取其中 COW 值高于阈值的样本,构建出新的少数类样本数据集,获得不平衡数据集 $D_{TL-COW-new}$ 。

3) 使用 SMOTE 算法生成新的样本。

3.2 模型训练

模型训练的 4 种方法包括 AdaBoost, RF, SVM 和 XGBoost。AdaBoost 是一种迭代算法,通过训练多个弱分类器并提升错误分类样本的权重,形成强分类器;RF 是由多个随机决策树组成的非参数监督学习方法,聚合所有树的预测结果对数据集进行分类;SVM 算法通过寻找最优超平面将不同类别的数据分开,将样本映射到高维空间,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中;XGBoost 是一种基于梯度提升框架的集成学习算法,能够快速处理大规模数据集,常用于机器学习训练模型。

本文中运用这 4 种训练方法与 SMOTE 不同改进方法结合,比较其对交通事故数据的学习能力和判断能力。

4 算法实践与结果分析

4.1 算法实践

数据集分为训练数据集和测试数据集,比例为 70% 和 30%。由于数据集数据总量较大,将总数据集随机分成 20 个有效数据集进行数据处理训练,每组数据集的有效数据约为 1 万条。对于不同训练模型,SMOTE 相关改进方法进行过采样生成的训练数据集不平衡比对训练结果有一定影响。本文设置训练数据集的不平衡比为 1:4,控制比例变量,使得生成数据集更具有有一般性,有助于实验结果分析和

讨论。

利用不同 SMOTE 改进方法进行数据改进,生成比例为 1 : 4 的平衡数据集,通过 t - 分布随机邻域嵌入数据降维方法 (t - Distributed Stochastic Neighbor Embedding, t - SNE), 创建缩小的特征空间,将高维数据转化至低维空间。由 ASMOTE 方法和 ISMOTE 方法生成训练集的降维结果如图 1 所示。

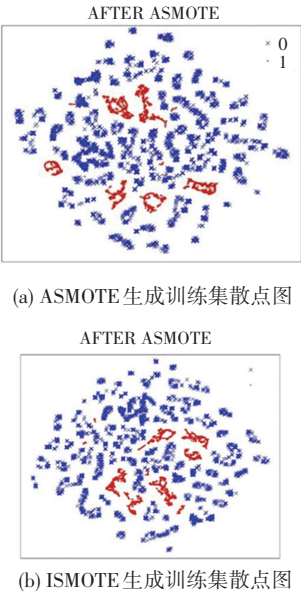


图 1 SMOTE 改进算法生成训练集数据示意图

Fig. 1 Diagrams of the training set generated by the improved SMOTE algorithms

训练数据用 SMOTE 改进算法过采样后,代入不同训练模型即 AdaBoost、RF、SVM、XGBoost 进行训练,并与基于原始不平衡训练数据集训练的模型比较,采用召回率 ($Recall$)、 $F1$ 分数 ($F1$)、精确率 (ACC)、曲线下面积 (AUC) 4 个指标量化评估实验效果。由于本实验涉及多次试验,且每次试验中数据集差别不大。为简洁起见,展示第一次试验数据集 (训练集和数据集数据) 的基本情况见表 2。

表 2 第一次试验数据集基本情况		
Table 2 Basic information of sets in the first experiment		
数据集	数据量	不平衡比
原始训练集	9 969	1 : 142
ASMOTE 生成训练集	12 373	1 : 4
ISMOTE 生成训练集	12 348	1 : 4
测试集	3 031	1 : 10

为了减少实验误差和实验结果的偶然性,本文进行了 10 次重复实验,并以平均水平代表各模型的平均预测水平,测试集预测结果见表 3。

表 3 测试集训练效果				
Table 3 Training effect of the test sets				
重采样算法	AUC	ACC	Recall	F1
原始数据+SVM	0.550 0	0.973 8	0.100 0	0.181 8
原始数据+AdaBoost	0.916 7	0.995 1	0.833 3	0.909 1
原始数据+RF	0.883 3	0.993 2	0.766 7	0.867 9
原始数据+XGBoost	0.835 1	0.982 9	0.307 7	0.013 0
ASMOTE+SVM	0.866 7	0.992 2	0.733 3	0.846 2
ASMOTE+AdaBoost	0.850 0	0.991 3	0.700 0	0.823 5
ASMOTE+RF	0.950 0	0.997 1	0.900 0	0.947 4
ASMOTE+XGBoost	0.999 3	0.994 3	0.984 8	0.970 1
ISMOTE+SVM	0.833 3	0.990 3	0.666 7	0.800 0
ISMOTE+AdaBoost	0.950 0	0.997 1	0.900 0	0.947 4
ISMOTE+RF	0.966 7	0.998 1	0.933 3	0.965 5
ISMOTE+XGBoost	1.000 0	0.993 3	0.982 1	0.964 8

4.2 实验结果分析

根据评估指标统计结果可见,基于 SMOTE 改进方法生成数据集的训练模型均一定程度上优于基于原数据的训练模型。就精确率 (ACC) 而言,由于测试集中少数类样本与多数类样本的不平衡比率较大,因此精确率主要体现判断多数类样本的准确率,不同算法在判断多数类样本的准确率上差距不大;就曲线下面积 (AUC) 评价指标而言,ASMOTE+XGBoost 与 ISMOTE+RF 均效果良好,相较于只使用原始数据的模型训练方法,分别提升了 20%、9%;就召回率 ($Recall$) 而言,由于 AdaBoos 训练的数据模型对不平衡数据的处理能力较高,ASMOTE+AdaBoost 相较于只使用原始数据的模型训练方法降低了 16%,ISMOTE+AdaBoost 相较于只使用原始数据的模型训练方法提升了 8%;就 $F1$ 分数而言,ISMOTE+RF 训练模型的稳定性较强,相较于使用原始数据的模型训练方法、ASMOTE+RF 分别提升了 6.2%、1.9%。

综上,通过对不同算法模型的比对和分析,本文提出的 ASMOTE+XGBoost、ISMOTE+RF 对交通事故的判别均具有较强的稳定性和较高的准确率,尤其对少数类样本的判别显著优于其他涉及模型,交通数据集预测效果较好。同时,实验结果表明,基于深度学习和机器学习的分类模型对交通事故预测判别等非线性案例处理更加高效。

5 研究方向与展望

SMOTE 方法在处理不平衡数据时基于线性插值表现出良好的优势,具有较强的实操性。然而随着大数据的普及和应用,在面临不同类型不平衡数据时,仍需进一步研究,利用 SMOTE 与其他技术相结合提升不平衡数据的建模精度。本文采用的 SMOTE 方法及其改进方法在实践过程中仍然存在

不足,可以在以下3个方面继续改进和完善。

(1)探究SMOTE及其改进方法对多源数据集的改进效果。本文仅选取特定路段的交通数据集进行效果实践,时空特征具有地域局限性,探究多源数据集的训练效果将是进一步研究的方向;

(2)调整测试集的数据比例,进一步检验模型效果。实验中发现测试集中少数类样本占比较低,导致量化评估指标中多数类样本被放大,少数类样本影响小,影响模型实用性。后续可提高少数类样本判别正确率权重,使实验结果更加贴近实际;

(3)数据级改进与算法级改进结合,探究不平衡数据改进的新方向。本文分别聚焦于SMOTE方法的数据级改进和算法级改进,后续研究可结合二者对SMOTE算法改进的优势,如通过成本敏感学习、集成学习等进一步扩大SMOTE算法的应用场景。

6 结束语

本文聚焦交通事故预测这一与社会文明发展密切相关的问题,通过对机器学习、不平衡数据分类和特征提取等的研究,将数据生成模型和类型预测模型融为一体。实验结果表明,本文所提及的ISMOTE+RF、ASMOTE+XGBoost模型在交通事故预测中表现出较为优越的性能,相较于其他传统预测模型能够更加有效地避免对少数类样本的误分类,提升了预测的准确性和鲁棒性,模型贴合实际运用需求,应用效果较好,对预测交通事故风险,预防大型交通事故有较为显著的作用。

改进的合成过采样方法在提高模型预测效果方面起到了关键作用,为交通事故的风险预测和防范提供了更为可靠的技术支持,为后续的研究提供了一定经验和启示。未来的研究可以在现有模型基础上进一步优化数据采样策略,实现更加精准和及时的交通事故风险预测。

参考文献

- [1] CAI Q, ABDEL-ATY M, YUAN J, et al. Real-time crash prediction on expressways using deep generative models [J]. *Transportation Research Part C: Emerging Technologies*, 2020, 117(1): 102697.
- [2] LIU M, CHEN Y. Predicting real-time crash risk for urban expressways in China [J]. *Mathematical Problems in Engineering*, 2017, 2017(1): 6263726.
- [3] 李昂,韩萌,穆栋梁,等. 多类不平衡数据分类方法综述 [J]. *计算机应用研究*, 2022, 39(12): 3534-3545.
- [4] HOSSAIN M, ABDEL-ATY M, QUDDUS M A, et al. Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements [J]. *Accident Analysis and*

- Prevention*, 2019, 124: 66-84.
- [5] CHEN F, CHEN S R, MA X X. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data [J]. *Journal of Safety Research*, 2018, 65: 153-159.
- [6] LORD D, MANNERING F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives [J]. *Transportation Research Part A - Policy and Practice*, 2010, 44(5): 291-305.
- [7] SCHLÖGL M, STÜTZ R, LAAHA G, et al. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset [J]. *Accident Analysis and Prevention*, 2019, 127: 134-149.
- [8] ABOU ELASSAD D E, ABOU ELASSAD Z E, ED-DAHBI A M, et al. An advanced accident avoidance system based on imbalance-control ensemble and deep learning fusion design [J]. *Journal of Intelligent & Fuzzy Systems*, 2024, 46(2): 3659-3676.
- [9] 吴秋伶,刘孙俊,王杰,等. 基于SMOTE和GWO-XGBoost的变压器故障诊断研究[J/OL]. *计算机测量与控制*, 1-12 [2025-04-13]. <http://kns.cnki.net/kcms/detail/11.4762.TP.20250212.1016.016.html>.
- [10] PANDE A, DAS A, ABDEL-ATY M, et al. Estimation of real-time crash risk are all freeways created equal? [J]. *Transportation Research Record*, 2011 (2237): 60-66.
- [11] YU R J, ABDEL-ATY M. Utilizing support vector machine in real-time crash risk evaluation [J]. *Accident Analysis and Prevention*, 2013, 51: 252-259.
- [12] Schlögl M. A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach [J]. *Accident Analysis & Prevention*, 2020, 136: 105398.
- [13] BASSO F, BASSO L J, BRAVO F, et al. Real-time crash prediction in an urban expressway using disaggregated data [J]. *Transportation Research Part C - Emerging Technologies*, 2018, 86: 202-219.
- [14] 石洪波,陈雨文,陈鑫. SMOTE过采样及其改进算法研究综述 [J]. *智能系统学报*, 2019, 14(6): 1073-1083.
- [15] 王晓霞,李雷孝,林浩. SMOTE类算法研究综述[J]. *计算机科学与探索*, 2024, 18(5): 1135-1159.
- [16] 李辉,李光旭. 基于AHP的SMOTEBagging改进模型 [J]. *电子科技大学学报(社科版)*, 2018, 20(4): 40-46.
- [17] KOSOLWATTANA T, LIU C, HU R, et al. A self-inspected adaptive SMOTE algorithm for highly imbalanced data classification in healthcare [J]. *BioData Mining*, 2023, 16(1): 15.
- [18] ILEBERI E, SUN Y, WANG Z. Performance evaluation of machine learning methods for credit card fraud detection using smote and AdaBoost [J]. *IEEE Access*, 2021, 9: 165286-165294. DOI:10.1109/access.2021.3134330.
- [19] 据春华,陈冠宇,鲍福光. 基于kNN-Smote-LSTM的消费金融风险检测模型——以信用卡欺诈检测为例 [J]. *系统科学与数学*, 2021, 41(2): 481-498.
- [20] 朱莉,汪小豪,李豪,等. 不平衡样本下基于变异麻雀搜索算法和改进SMOTE的变压器故障诊断方法 [J]. *高电压技术*, 2023, 49(12): 4993-5001.
- [21] FENG M, WANG X, CAI B, et al. Hourly traffic crash prediction using environmental and electric vehicle big data [C]// *Proceedings of the Transportation Research Board 101th Annual Meeting*. Washington DC: TRB, 2022.