武汉理工大学学报(交通科学与工程版)
Journal of Wuhan University of Technology(Transportation Science & Engineering)
ISSN 2095-3844,CN 42-1824/U

《武汉理工大学学报(交通科学与工程版)》网络首发论文

题目：基于权重的欠采样提升算法识别激进驾驶员
作者：彭一川，李崇奕，王可，邢莹莹
收稿日期：2020-01-14
网络首发日期：2020-10-10
引用格式：彭一川，李崇奕，王可，邢莹莹．基于权重的欠采样提升算法识别激进驾驶员．武汉理工大学学报(交通科学与工程版).
https://kns.cnki.net/kcms/detail/42.1824.U.20201010.0849.008.html

基于权重的欠采样提升算法识别激进驾驶员

彭一川，李崇奕，王可，邢莹莹
(同济大学道路与交通工程教育部重点实验室 上海 201804)

摘要：机器学习算法广泛应用于危险驾驶行为和危险驾驶人的识别。由于危险驾驶行为或危险驾驶人在实际交通中的比例非常低，所以一般的机器学习算法更倾向学习如何识别正常样本而不是危险样本。而对于危险驾驶行为研究学者来说，危险样本才是研究的重点。文章旨在利用车辆轨迹数据以及不平衡类提升算法来识别驾驶员的跟驰行为(正常和激进)。首先，提出了一种碰撞风险的度量方法，即平均碰撞风险(Average Crash Risk, ACR)。该方法基于驾驶员在车辆跟随过程中的响应数据来计算车辆的平均碰撞风险。其次，基于平均碰撞风险，对每一位驾驶员的激进驾驶行为进行判断并且赋予标签。最后，将车间距数据进行离散傅里叶变换作为输入变量进行激进驾驶员识别算法的训练。使用高精度 NGSIM 数据集进行激进驾驶员识别，研究对象包括在美国 I-80 高速公路 HOV 车道上无换道干扰的 299 对跟驰车辆。此外，文章提出一种新的基于权重的欠采样提升算法 WUSBoost（Weight-based Under Sampling Boost），并与其它不平衡类提升算法 SMOTEBoost 和 RUSBoost，常规的提升算法 AdaBoost 和 XGBoost，以及 SMOTE+AdaBoost 和 RUS+AdaBoost 等算法进行比较。在所有算法中，WUSBoost 的性能最好，显示了其对于不平衡类数据有较好的识别能力。

关键字：车辆轨迹；激进驾驶识别；基于权重；欠采样提升算法；不平衡数据
中图分类号：U491.31

收稿时间：2020-01-14
作者简介：第一作者，彭一川，（1982-），男，博士，副教授，研究领域为道路交通安全。E-mail：yichuanpeng@tongji.edu.cn
通讯作者，王可（1986-），男，博士后，研究领域为驾驶行为与智能交通安全。E-mail: kew@tongji.edu.cn
基金项目：国家重点研发计划（2018YFB1201403），国家自然科学基金项目（71871165）资助

0 引　言

驾驶风险识别在驾驶安全与辅助驾驶系统中起着重要的作用。随着联网自动驾驶汽车和先进的驾驶员辅助系统(ADAS)的发展，通过人工智能手段识别激进型的驾驶员和防止撞车事件的可能性与需求越来越大。近年来，许多研究致力于使用安装多种传感器的车辆进行自然驾驶实验，通过实验数据进行驾驶行为识别[1-3]。比如，一些学者使用车载传感器进行自然驾驶数据的采集，以获取驾驶员实时的视频图片、油门信息、踏板制动、速度以及加速度等信息[4-7]。也有一些研究学者采用驾驶模拟器来监测在预设计的驾驶环境中的驾驶行为[8,9]。与前两种方法相比，部署在路边的视频监控可以以较低成本提供海量的交通环境数据和车辆轨迹数据[10]。从视频监控中提取的车辆轨迹数据由横向和纵向数据组成。Murphey 等人[11]采用了包括速度、加速度在内的纵向数据，将驾驶员分为激进型、温和型和冷静型。横向数据，即偏航率、与车道线的距离、侧向加速度也被用来识别驾驶行为[12]。这些学者的研究证明了，通过视频提取的车辆轨迹数据是可用于进行驾驶行为识别研究。

在如何衡量驾驶安全水平的问题上，以往的研究广泛地应用了一些碰撞风险指标，如碰撞时间(Time To Collision，TTC)，车头时距(Headway)和 DRAC（Deceleration Rate to Avoid a Crash）等。Mahmud 等人[13]比较了碰撞风险指标之间的优缺点，如 TTC 不能很好地处理车辆跟驰时相对速度为零的问题，并且需要假设前车和跟驰车辆的速度都是恒定的。由 Kitajima 等人首先提出的 MTC [14]测量的是在前车和跟车都突然减速的情况下的碰撞可能性。DSS（Difference of Space distance and Stopping distance）[15]和 PICUD（Potential Index for Collision with Urgent Deceleration）[16]进一步考虑跟车车辆减速的反应时间。TIDSS [17] 通过对某一时间段内 DSS 与危险阈值之间的差值进行积分，计算出车辆的综合碰撞风险。这些研究为进一步建立基于视频监控车辆轨迹的碰撞风险指标提供了参考，然而以往的研究对于风险指标阈值的确定仍然缺乏更为合理的方法。

在驾驶行为识别模型的研究中，机器学习算法一直受到青睐。如学者们常用神经网络[6,18]、隐形马尔科模型[19]或者支持向量机[20-22]进行驾驶行为识别。但在危险驾驶行为研究中，数据通常是不平衡的。比如，在现实交通中，激进型驾驶员的数量要比正常驾驶员少。大多数机器学习算法往往关注于数据量较大的正常驾驶行为数据标签。而激进型驾驶行为作为数据集中的少数，可能会被分类算法忽略。因此，在机器学习算法的训练中使用不平衡的数据可能会导致有偏差的结果，从而导致模型对激进驾驶员的识别性能下降。已有研究一般会在使用分类算法前对不平衡样本进行预采样，使其不平衡程度降低[23,24]。预采样的方法包括 SMOTE(synthetic minority oversampling technique)，随机欠采样等。SMOTE(Synthetic Minority oversampling Technique)是由 Chawla 等人[25]提出的，通过在特征空间中合成少数类的更多实例，进而扩展少数类的决策区域和平衡类别比例。然而，据作者所知，在危险驾驶行为识别领域还没有使用不平衡类提升算法的研究发表。诸如 SMOTEBoost [26]、RUSBoost [27]等不平衡类提升算法旨在将提升算法与采样方法相结合，在每一轮的提升学习中都使用采样来改变样本分布。不平衡类提升算法在多个数据集上的测试得到了比单独使用采样方法更好的结果[26,27]。

采用视频监控中提取的 NGSIM 车辆轨迹数据，通过一种新的碰撞风险度量方法进行激进驾驶员的标注。接着基于带有标签的不平衡数据，致力于通过不平衡算法研究提高激进型驾驶员识别效果。最后对所有算法的识别性能指标进行了讨论。

1　数据描述

NGSIM 是目前研究使用最多的车辆轨迹公共数据集。该数据集使用摄像机采集交通流

视频，并且通过计算机视觉技术提取车辆轨迹数据。NGSIM 的一部分数据来自美国加州 I-80 高速公路。该路段长度约 500 米，包含 6 条车道，包括一个高占用率车辆(High Occupancy Vehicle, HOV)车道。数据采集时间为 2005 年 4 月 13 日下午 4:00—4:15，以及下午 5:00—5:30，共 45 分钟。

虽然 NGSIM 数据在学术研究中被广泛应用，尤其是在交通流理论领域，但其准确性近年来受到质疑[28-30]。NGSIM 数据集的测量误差远远超出了可以忽略的范围，部分原因是因为低分辨率相机和视频图像对车辆的错误跟踪。Montanino 和 Punzo[28]删除了异常和噪声，重建了部分 I-80 数据集(下午 4:00—4:15)，这比原来的 NGSIM 数据集精度有了显著的改进。

实验使用 Montanino 和 Punzo 的数据集进行激进型驾驶员识别，重点研究了在 HOV 车道上没有换道干扰的 299 对前车-跟车跟驰车辆(Leading-following Vehivle Pair, LVP)。每对前车-跟车跟驰车辆的跟驰时间至少持续 10 秒，通常超过 20 秒(如图 1 所示)。数据中每 0.1 秒记录下前车和跟车的速度、加减速率和间距。

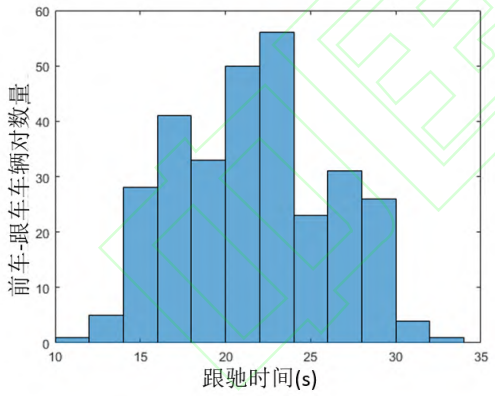

图 1 跟车时间直方图

2 建模方法

建模方法分为五部分介绍。第一部分主要阐述如何使用 NGSIM 中提取的信息来衡量每辆车的碰撞风险方法。第二部分讨论了如何确定 ACR 阈值来标注激进型驾驶员。第三部分阐释了离散傅里叶变换(Discrete Fourier Transform, DFT)方法，该方法将给定的时间序列转换为频域内信号幅值，从而揭示隐藏在车辆轨迹数据中的行驶特性。第四部分提出一种新的不平衡类提升强算法。最后，介绍用来衡量提升算法能力的性能指标。

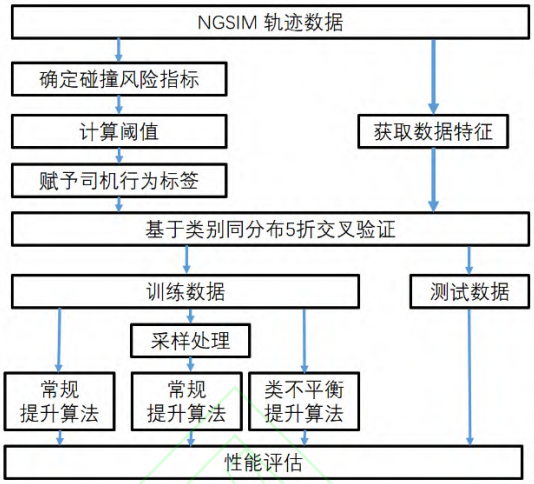

图 2 建模方法框架

2.1 碰撞风险指标

文章提出了一个新的碰撞风险指标，平均事故风险(ACR)来衡量驾驶员的激进性。对于每辆车，t 时刻的碰撞风险(CR)根据 DSS 计算:

$$\mathrm{DSS}(t)=\frac{v_l^2(t)-v_f^2(t)}{2\mu g}+d(t)-\tau v_f(t) \quad (1)$$

式中：v_l 、v_f分别是前车和跟车的速度；μ 是轮胎与地面的摩擦系数，设为 0.7; g 是重力加速度，设为 9.8m/s^2; d 是车间距离;τ 是驾驶员的反应时间，当车辆加速时设为 1.5s，车辆减速时设为 0.75s。

当 DSS >为 0 时，表示跟车车辆有足够的时间减速避免碰撞，因此，碰撞风险为 0。当 DSS≤0 时，跟车车辆存在潜在的碰撞风险，碰撞风险用 DSS 的绝对值除以跟车车辆的速度来衡量。

$$\mathrm{CR}(t)=\begin{cases}0, & DSS(t)>0\\ \dfrac{|DSS(t)|}{v_f(t)}, & DSS(t)\le 0\end{cases} \quad (2)$$

2.2 平均碰撞风险阈值(ACR 阈值)

为了衡量每个驾驶员在整个跟车过程中的整体驾驶激进性，平均碰撞风险计算如下:

$$\mathrm{ACR}=\frac{1}{T}\sum_{t=0}^{T}CR(t)*\Delta t \quad (3)$$

式中：T是跟车时间；Δt 是样本间隔时间，为 0.1s。

当驾驶员的 ACR 超过一定的阈值时，便赋予该名驾驶员的驾驶行为标签。然而，在以

往的研究中并没有经验或理论的阈值，因此，通过 K-means 聚类算法来寻找正常驾驶员和激进型驾驶员之间的边界。

给定一组观测值 $(x_1, x_2, \cdots, x_n)$，其中X_i是一个 d 维实向量。K-means 聚类目的是将 n 个观测值分成 k 簇$\{C_1, C_2, \cdots, C_k\}$，从而使簇内方差和最小化。K-means 的目标函数如下:

$$\text{Min E} = \sum_{i=1}^{k} \sum_{x=C_i} ||x - \mu_i||_2^2 \qquad (4)$$

式中：$\mu_i = \frac{1}{|C_i|} \sum_{x=C_i} x$表示 C_i簇的向量均值

通过 Matlab 内置函数 kmeans 将 299 对驾驶员数据分为两类:正常驾驶行为簇和激进型驾驶行为簇。kmeans 函数使用平方欧氏距离度量，并且通过启发式方法来寻找 k-means 类簇的中心。

2.3 离散傅里叶变换

由于每个 LVP 的跟车过程时间长度不同，所以不能直接用每辆车的速度和加速度的时间序列来识别驾驶员的驾驶行为。离散傅里叶变换(DFT)在驾驶行为研究中被多次应用，将驾驶特征的时间序列在频域内转换为信号幅度。

给定时间序列$(x_1, x_2, \ldots, x_N)$的 DFT 定义为 N 个复数$(DFT_0, DFT_1, \ldots, DFT_{N-1})$：

$$DFT_K = \sum_{n=0}^{N-1} x_n e^{(-\frac{2\pi i}{N} kn)} \qquad (5)$$

式中：i是虚数单位。

使用每个时间序列的前 15 个 DFT 系数的实数部分作为机器学习的直接输入，其代表了时间序列的低频区信号强度，高频区的信号作为噪声被移除。

2.4 不平衡类提升算法

在实际交通中，激进型驾驶员的比例远远小于普通驾驶员的比例。因此，数据集通常以激进型驾驶员作为少数类进行不平衡处理。Chawla 等人[26]提出了一种 SMOTEBoost 算法，该算法结合了合成的少数过采样技术(SMOTE)和标准的提升过程。标准的提升算法对所有错误分类的例子给出了相同的权重，通常在后续的迭代中大多数类会依旧占有很大的比例。通过在每一轮提升中引入 SMOTE 采样，SMOTEBoost 算法能逐渐增加少数类样本的数量。与 SMOTEBoost 相似，Seiffert 等人[27]提出了将随机欠采样 RUS 和 AdaBoost 相结合的 RUSBoost 算法。在每个提升的迭代中，使用 RUS 代替 SMOTE 用来平衡数据集。

文章提出一种新的基于权重的欠采样提升算法(WUSBoost)，其对 RUSBoost[27,31]进行改进，在每一轮提升迭代中对多数类样本进行聚类，并基于样本权重进行随机抽样，只保留部分多数类样本进入下一轮提升学习。

假设拥有 n 个样本训练集 S（样本标签为 K=2 类，其中数量较多一类数据以下称为多数类，反之，称为少数类），M 个弱分类器，设定聚类个数为 C_num，欠采样的倍率为 p(p∈(0,1))，算法步骤如下表 1 所示。

表 1　WUSBoost 算法流程

1、初始化样本权重为$w_1(\text{i}) = 1/n$ ，i =1,2,⋯,n

2、进行 M 次迭代操作，第 m 次迭代操作如下(m=1,2,⋯,M):

2.1、使用 K-均值聚类算法将多数类样本分类 C_num 个簇；将每个簇内样本的权重作为轮盘赌的生成概率，进行有放回抽样，抽样后得到的新样本数量为原来簇内样本数量的 p(p∈(0,1))倍；合并所有簇产生的新多数类样本以及少数类样本，得到新样本集合s'_m，新样本个数为n'_m以及对应的样本权重w'_m；

2.2、基于样本集合s'_m和样本权重分布w'_m，进行第 m 号分类器$T^{(m)}(x)$的训练；

2.3、获得该分类器的样本类别概率估计：

$$p_k^{(m)}(x) = Prob_w(c = k|x), k = 1, \ldots, K$$

2.4、令：

$$\text{h}_k^{(m)}(x) = (K-1)\left(logp_k^{(m)}(x) - \frac{1}{K}\sum_k logp_k^{(m)}(x)\right), \text{k} = 1, \ldots, \text{K}$$

2.5、更新样本权重：

$$w_i = w_i * \exp\left(-\frac{K-1}{K} y_i^T log p_k^{(m)}(x_i)\right), \mathrm{i} = 1, \ldots, n_m$$

2.6、样本权重标准化：

$$w_i = w_i / \sum_i^n w_i \quad , \quad \mathrm{i} = 1, \ldots, n_m$$

3、得到强分类器，输出

$$\mathrm{C(x)} = \mathrm{sign}(\arg\max \sum_{m=1}^{M} \mathrm{h}_k^{(m)}(x))$$

除上述三种不平衡类提升算法外，还对其他 6 种算法进行了测试。所有算法见表 2。

表 2 测试算法列表

算法名称	提升算法前的预采样
AdaBoost	无
XGBoost	无
SMOTE+ AdaBoost	SMOTE
SMOTE+ XGBoost	SMOTE
RUS+ AdaBoost	随机欠采样
RUS+ XGBoost	随机欠采样
SMOTEBoost	无
RUSBoost	无
CUSBoost	无

2.5 性能评估

提升算法的性能可以通过其利用车辆轨迹数据识别激进型驾驶员的能力来反应。评估识别模型主要使用以下三个重要的性能指标: 精确率、召回率和 F1 评分。

精确率定义如下:

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \qquad (6)$$

式中：TP 是被正确识别的激进型驾驶员的数量，FP 是被错误地识别为激进型驾驶员的普通驾驶员的数量。

召回率定义如下:

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \qquad (7)$$

式中：FN 是被错误地识别为正常驾驶员的激进驾驶员的数量。

F1 分数是精确率和召回率的调和平均值。F1 分数在 1 处达到最佳值(精确度和召回率均为 1)，在 0 处达到最差值。F1 公式定义如下：

$$\mathrm{F1} = \frac{2*\mathrm{TP}*\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \qquad (8)$$

ROC 曲线(receiver operating characteristic curve)是表示 0 ~ 1 之间不同候选阈值的假阳性率与真阳性率的曲线图。类似地，精确率-召回率曲线是不同阈值下的准确率和召回率的曲线。一般情况下，当每个类别的观察值大致相等时，应使用 ROC 曲线。当存在类不平衡时，应该使用精确率-召回率曲线。因为不平衡数据集的 ROC 曲线可能具有欺骗性，并导致对模型性能[32]的错误解释。因此，使用精确率-召回率曲线下的面积(Area Under Precision-Recall Curve, AUPRC)来比较算法的性能，通过测量整个精确率-召回率曲线下的二维区域得到的面积值作为算法不平衡处理性能指标。

K折分层交叉验证被广泛用于评价分类算法的性能，用来防止可能产生过拟合问题。交叉验证方案将 299 名驾驶员随机分为 K=5 个样本数相近的子集，并保证每个子集中。每次验证过程使用四个子集进行训练，并使用剩下来的一个子集来评估训练模型的性能。为尽量去除样本分配时的随机性，将五折分层交叉验证重复 5 次，总共产生了 25 组测试集合训练集，并将 25 次的少数类（在此案例中为激进驾驶员）精确率、召回率、F1 和 AUPRC 的平均值代表算法的性能。

3 结果分析

3.1 平均碰撞风险阈值(ACR 阈值)

依据上述建模方法，利用 NGSIM 轨迹数据可以计算出 LVP 中每辆跟车车辆的平均碰

撞风险(ACR)。如图 3 显示了 299 名驾驶员的平均撞车风险直方图。ACR 反映了跟车车辆中驾驶员的驾驶激进程度：35.7%的驾驶员 ACR=0，这意味着他们在跟驰过程中从未出现负的 DSS 值，总是保留足够的车间距和反应时间来避免碰撞。65.8%的驾驶员 ACR < 0.1，这表明在汽车跟随过程中只出现偶然和暂时的碰撞风险。

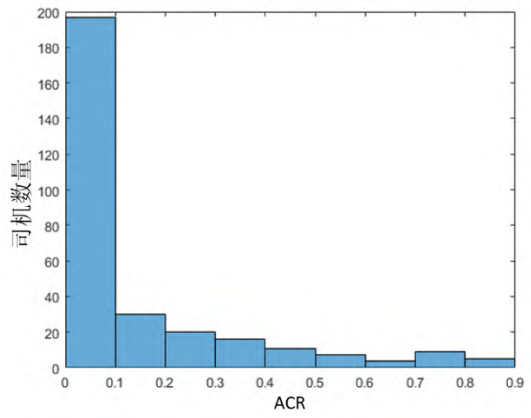

图 3　ACR 阈值直方图

基于 K-means 聚类结果分析，可以认为 ACR 值高于 0.14 的驾驶员可被认为是激进型驾驶员，ACR 值低于 0.14 为正常驾驶员。在 299 辆跟车驾驶员中，43 位被标定为激进驾驶员，占比 14.3%；256 位被标定为正常驾驶员，占比 85.6%。因为正常驾驶员的数量远远大于激进驾驶员的数量，所以驾驶员类别数据是不平衡的。

3.2 算法结果

基于上述驾驶员标定结果，使用跟驰车间距的离散傅里叶系数作为输入变量，训练 WUSBoost 等九种算法来识别激进驾驶员。通过最优化调参后，每种算法的平均精确率、召回率、f1 评分和 AUPRC 见表 3。

表 3 显示，WUSBoost 的召回率较高，为 0.944，仅低于 RUSBoost 的 0.962，这表示对于所有被识别为激进驾驶员的人中，只有约 5%是“假警报”。WUSBoost 的精确率要高于 RUSBoost。RUSBoost 对激进驾驶员的准确率只有 0.588，这意味着被识别为激进驾驶员的人中有 41.2%的为误判。XGBoost 与 SMOTE+XGBoost 给出了最高的 F1 值，0.897 与 0.903，这意味着这两个算法能保证精确率与召回率同时在较好的水平。由于对于同一个算法，精确率和召回率为此消彼长的关系，在提升一个指标的同时必然会带来另一个指标的下降。因此，可以使用 AUPRC 指标综合衡量精确率和召回率在不同的组合下是否能一直保持较高的数值。WUSBoost 总体上生成的结果最好,其 AUPRC 是 0.930，为所有算法中最高。结果表明，单独使用车间距作为输入变量依然能产生很好的识别效果。

表 3 提升算法性能表

算法	准确率	召回率	F-1 值	AUPRC
AdaBoost	0.832	0.768	0.786	0.852
SMOTE + AdaBoost	0.845	0.824	0.825	0.869
RUS + AdaBoost	0.681	0.901	0.774	0.820
XGboost	0.910	0.894	0.897	0.917
SMOTE+XGboost	0.887	0.93	0.903	0.902
RUS+XGboost	0.823	0.917	0.861	0.890
SMOTEBoost	0.799	0.856	0.818	0.895
RUSBoost	0.588	0.962	0.722	0.851
WUSBoost	0.659	0.944	0.761	**0.930**

图 4 展示了 25 次交叉验证中的一次训练结果。随着各种提升算法的迭代，少数类的 AUPRC 大多逐渐增长。可以看出使用了随机欠采样的 RUS+AdaBoost、RUS+XGBoos，和 RUSBoost 均有较低的 AUPRC。这可能是由于随机欠采样一次性剔除了大量多数类样本造成的。不平衡比例越大，剔除的多数类样本越多，导致了数据信息严重流失，影响了模型的训练效果。在这一次训练结果中，WUSBoost 和 SMOTE+XGBoost 的 AUPRC 最高，但 WUSBoost 在迭代进行了 5 轮左右时已经达到了很高的 AUPRC 值，而

SMOTE+XGBoost 在迭代了 30 次左右时才追赶上 WUSBoost 的水平。

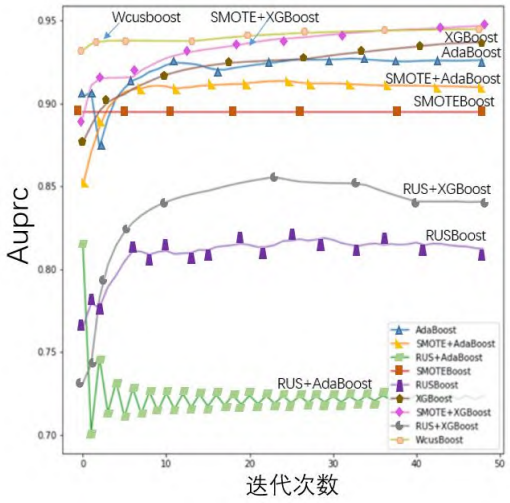

图 4 每轮提升学习后的 AUPRC 变化

本数据中激进驾驶员的比例为 14.4%，两类驾驶员的比例约为 6:1，不平衡比例并不是特别严重。下一步，提高 ACR 阈值到 0.28，减少激进驾驶员的数量到 19 位，加大不平衡比例到 14:1。表 4 展示了所有算法在更不平衡的数据下的性能表现。

WUSBoost 仍然得到了最高的 AUPRC，0.911。由于数据不平衡比例加大，XGBoost、SMOTE+XGBoost 等在上一个数据中表现良好的算法性能大幅下降。XGBoost 的 AUPRC 从上一个数据中的 0.917 下降到 0.843，而 SMOTE+XGBoost 的 AUPRC 从 0.902 下降到 0.837。

表 4 提升算法性能表（加大不平衡比例的数据）

算法	准确率	召回率	F-1 值	AUPRC
AdaBoost	0.853	0.847	0.832	0.857
SMOTE + AdaBoost	0.869	0.940	0.888	0.813
RUS + AdaBoost	0.613	0.991	0.743	0.806
XGboost	0.869	0.863	0.846	0.843
SMOTE+XGboost	0.877	0.993	0.930	0.837
RUS+XGboost	0.611	0.997	0.741	0.851
SMOTEBoost	0.867	0.900	0.867	0.872
RUSBoost	0.798	0.960	0.854	0.903
WUSBoost	0.841	0.970	0.884	**0.911**

根据算法测试结果，发现不平衡类提升算法 WUSBoost 在处理不平衡数据时性能优于常规提升算法和其他不平衡类提升算法。数据不平衡度越高，不平衡类提升算法的优势越明显。当激进型驾驶员约占样本量的 14.4%时，WUSBoost 和表现最好的常规提升算法 XGBoost 之间的性能差距很小(0.930 比 0.917)。相比之下，当提高 ACR 阈值到 0.28，识别只占样本量 6.4%的激进类驾驶员时，WUSBoost 的 AUPRC 从 0.930 下降到 0.911，而表现最好的常规算法 AdaBoost 只有 AUPRC=0.857。另外，使用预采样方法后的常规提升算法往往结果比不使用预采样更糟。表 2 中，SMOTE+XGBoost 与 RUS+XGBoost 的 AUPRC 均低于 XGBoost。表 3 中，SMOTE+AdaBoost 与 RUS+AdaBoost 的 AUPRC 均低于 AdaBoost。这说明在提升算法之前采用采样方法并不能保证一定提高算法识别能力。一个可能的原因是，在提升算法迭代训练前对训练数据进行一次性预采样，可能会造成训练集和测试集的数据分布的偏差。而测试集中类别比例应与实际应用中的真实数据类别比例相近，是不能进行预采样的。这种训练集和测试集的差异会削弱训练模型对测试数据的识别能力。

4 结 论

文章中利用不平衡类提升算法和重构的 NGSIM 轨迹数据，探讨了识别激进型驾驶员的可能性。为了标注激进驾驶员，提出了碰撞

风险度量指标——平均碰撞风险(Average Crash Risk, ACR)。ACR 根据驾驶员在跟车过程中的驾驶反应来区分激进型的驾驶员和其他驾驶员。相对于专家的主观判断、问卷调查、基于速度/加速度/车轮转向的聚类等其他标签方法，碰撞风险指标更客观表示了驾驶员跟驰行为的危险程度。

提出了一种基于权重的不平衡类提升算法 WUSBoost，与其他不平衡类提升算法 SMOTEBoost 和 RUSBoost 相比具有更好的识别不平衡数据的能力。数据不平衡度越高，不平衡类提升算法与 XGBoost 等常规提升算法相比优势越大。WUSBoost 是采样方法和 AdaBoost 算法的结合，因此当使用更先进的提升算法替代 AdaBoost 时，WUSBoost 的性能可能会进一步提高。预采样方法并不一定会提高常规提升算法的表现，实验中发现使用了 SMOTE、RUS 等预采样后的训练数据集往往会影响 AdaBoost、XGBoost 在训练数据集上的性能。

当使用车间距的离散傅里叶系数作为输入时，激进驾驶员识别模型可以达到较高的准确率和召回率。这意味着即使不使用车内传感器获取驾驶员操控行为，比如方向盘转角、油门踏板等，仅通过路侧监控视频提取的车间距数据仍能较好的识别激进的驾驶员，从而起到提前预警，规避风险的作用。

参考文献

[1] KLUGER R, SMITH B, PARK H, et al. Identification of safety-critical events using kinematic vehicle data and the discrete fourier transform[J]. Accident Analysis and Prevention, 2016, 96: 162-168.

[2] SCHORR J, HAMDAR S H, SILVERSTEIN C. Measuring the safety impact of road infrastructure systems on driver behavior: vehicle instrumentation and real-world driving experiment[J]. Journal of Intelligent Transportation Systems,2017,21(5): 364-374.

[3] SUN W, ZHANG X, PEETA S, et al. A real-time fatigue driving recognition method incorporating contextual features and two fusion levels[J]. IEEE Transactions on Intelligent Transportation Systems, 2017,18(12): 3408-3420.

[4] LIU T, YANG Y, HUANG G B, et al. Driver distraction detection using semi-supervised machine learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2016,17(4): 1108-1120.

[5] CHANDRASIRI N P, NAWA K, ISHII A. Driving skill classification in curve driving scenes using machine learning[J]. Journal of Modern Transportation,2016,24(3): 196-206.

[6] MOLCHANOV P, GUPTA S, KIM K, et al. Multi-sensor system for driver's hand-gesture recognition[C]. IEEE International Conference and Workshops on Auto-matic Face and Gesture Recognition, Ljubljana, Slovenia, 2015.

[7] WU M, ZHANG S, Dong Y. A novel model based driving behavior recognition system using motion sensors[J]. Sensors, 2016,16(10):1746.

[8] FERNANDEZ S, ITO T. Driver classification for intelligent transportation systems using fuzzy logic[C]. IEEE International Conference on Intelligent Transportation Systems, Rio de Janeiro, Brazil, 2016.

[9] WANG W, XI J. A rapid pattern-recognition method for driving styles using clustering-based support vector machines[C]. American Control Conference, Boston, USA, 2016.

[10] YAN C, COENEN F, YUE Y, et al. Video-based classification of driving behavior using a hierarchal classification system with multiple features[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(5): 1-33.

[11] MURPHEY Y L, MILTON R,KILIARIS L.Driver's style classification using jerk analysis[C]. IEEE Workshop on

Computational Intelligence in Vehicles and Vehicular Systems, Nashville, USA, 2009.

[12] HIGGS B, ABBAS M. Segmentation and clustering of car-following behavior: recognition of driving patterns[J]. IEEE Transactions on Intelligent Transportation Systems, 2015,16(1):81-90.

[13] MAHMUD S M S, FERREIRA L, HOQUE M S, et al. Application of proximal surrogate indicators for safety evaluation: a review of recent developments and research needs[J]. IATSS Research, 2017, 41(4): 153-163.

[14] KITAJIMA S, TAKATORI O, ENOKID-A S, et al. Estimation of driver's dangerous states of rear-end collision based on driver video recorder data and ordinary driving data[C]. Proceedings of automotive engineers of Japan, 2009.

[15] Japan Society of Traffic Engineers. Traffic engineering handbook[M]. 2005.

[16] UNO N, IIDA Y, ITSUBO S, et al. A microscopic analysis of traffic conflict caused by lane-changing vehicle at weaving section[C]. Proceedings of the 13th Mini EURO Conference on Handling Uncertainty in the Analysis of Traffic and Transportation Systems, 2003.

[17] OKAMURA M, FUKUDA A, MORITA H, et al. Impact evaluation of a driving support system on traffic flow by micro-scopic traffic simulation[J]. Advances in Transportation Studies, 2011, Special Iss-ue: 99-102.

[18] WANG H, ZHANG C, SHI T, et al. Real-time EEG-based detection of fatigue driving danger for accident prediction[J]. International Journal of Neural Systems, 2015, 25(2): 1550002.

[19] WANG W, XI J, ZHAO D. Driving style analysis using primitive driving patterns with bayesian nonparametric approaches[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(8): 2986-2998.

[20] SUN R, HAN K, HU J, et al. Integrated solution for anomalous driving detection based on BeiDou/GPS/IMU measurement[J]. Transportation Research Part C, 2016, 69: 193-207.

[21] CHEN Z, WU C, HUANG Z, et al. Dangerous driving behavior detection using video-extracted vehicle trajectory histograms[J]. Journal of Intelligent Transportation Systems, 2017, 21(5): 409-421.

[22] XUE Q, WANG K, LU J, et al. Rapid driving style recognition in car-following using machine learning and vehicle trajectory data[J]. Journal of Advanced Transportation. 2019: 1-11.

[23] SHI X, WONG Y D, LI M Z F, et al. A feature learning approach based on XGBoost for driving assessment and risk prediction[J]. Accident Analysis & Prevention, 2019,129: 170-179.

[24] 郑文昌，陈淑燕，王宣强.面向不平衡数据集的SMOTE-SVM交通事件检测算法[J]. 武汉理工大学学报，2012，11:64-68.

[25] 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(1):321-357.

[26] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia , 2003.

[27] SEIFFERT C, KHOSHGOFTAAR T, HULSE J V, et al. Rusboost: A hybrid approach to alleviating class imbalance[J], IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2010, 40(1): 185-197.

[28] MONTANINO M, PUNZO V. Trajectory data reconstruction and simulation-based validation against macroscopic traffic

patterns[J]. Transportation Research Part B, 2015,80: 82-106.

[29] PUNZO V, BORZACCHIELLO M T, CIUFFO B. On the assessment of vehicle trajectory data accuracy and application to the Next Generation Simulation (NGSIM) program data[J]. Transportation Research Part C, 2011, 19(6): 1243-1262.

[30] COIFMAN B, LI L. A critical evaluation of the next generation simulation (NGSIM) vehicle trajectory dataset[J]. Transportation Research Part B , 2017, 105: 362-377.

[31] HASTIE T, ROSSET S, ZHU J, et al. Multi-class AdaBoost[J]. Statistics & Its Interface, 2006, 2(3): 349-360.

[32] SAITO T, REHMSMEIER M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets[J]. PLoS One ,2015, 10(3):0118432.

Aggressive Driver Recognition using Weight-based under Sampling Boosting Algorithm

PENG Yichuan LI Chongyi WANG Ke XING Yingying

(The Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract：Machine learning algorithms are wildly applied in the recognition of risky driving behavior and dangerous drivers. Since the proportion of risky behavior or drivers in real traffic is very low, common machine learning algorithms prone to better recognize normal sample rather than risky sample, which is our real interest. This paper aims to use vehicle trajectory data and imbalanced class boosting algorithms to identify driver's driving style (normal vs. aggressive). First, a surrogate measurement of collision risk, called Average Crash Risk (ACR), is proposed to calculate vehicle's crash risk based on how driver response in car-following process. Then, the driver's driving style is determined by his/her average crash risk. Finally, the discrete Fourier transform of the distance data is used as the input to train the aggressive driver recognition algorithm. This paper uses reconstructed NGSIM dataset for aggressive driver identification and focuses on 299 leader-follower vehicle pairs on I-80 HOV lane that was not interrupted by lane-changing. This paper proposes a new weight-based under-sampling algorithm named WUSBoost (weight-based Under Sampling Boost). The new algorithm is compared with other unbalanced sampling algorithms (e.g. SMOTEBoost and RUSBoost), conventional Boosting algorithms (e.g. AdaBoost and XGBoost), and Boost +AdaBoost and RUS+AdaBoost, etc. The results show WUSBoost has the best performance of its kinds, indicating its better ability to recognize unbalanced data.

Key words: vehicle trajectory; aggressive driver recognition; weight-based；under sampling boost algorithm; imbalanced data