文章编号：0253-374X(2020)01-0051-09

DOI：10.11908/j.issn.0253-374x.19147

基于自然驾驶数据的危险事件识别方法

王雪松，徐晓妍

（同济大学 道路与交通工程教育部重点实验室，上海 201804）

摘要：利用阈值法从自然驾驶数据中识别可能的危险事件，再采用随机森林模型和支持向量机模型深度筛选，克服了阈值法误报率过高的缺陷。基于上海自然驾驶数据，建立提取危险事件的阈值标准，从原始数据中识别出 3 623 起可能的危险事件；利用随机森林模型筛选出重要特征作为输入变量，训练机器学习模型，对测试集进行预测。结果表明，起到关键作用的变量有：纵向加速度的最小值和均值、与前车距离的最小值以及车速的标准差。相比随机森林模型，支持向量机模型预测效果更优，在控制漏报率的同时，可过滤 85.9％的无效事件。

关键词：驾驶行为；自然驾驶研究；危险事件；机器学习

中图分类号：U491　　**文献标志码**：A

Detection of Safety-critical Events Based on Naturalistic Driving Data

WANG Xuesong, XU Xiaoyan

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

Abstract: Possible safety-critical events (SCEs) were identified from the naturalistic driving data using a threshold method. Random forests (RF) and support vector machine (SVM) models were employed to further screen the possible events, overcoming the defect of a high false positive rate while applying threshold methods solely. A set of threshold criteria was established and 3 623 possible SCEs were extracted from the naturalistic driving data in Shanghai. The RF method was adopted to select the important features as input variables. The RF and SVM models were trained and tested respectively on the same dataset. The results indicate that: the mean and minimum value of longitudinal acceleration, the minimum value of the distance from the leading vehicle and the standard deviation of the speed of the subject vehicle can effectively determine whether the possible events are valid or not. Compared with RF, SVM performs better in prediction, that is, filtering 85.9% invalid events and controlling false negative rate simultaneously.

Key words: driving behavior; naturalistic driving study; safety-critical event; machine learning

根据世界卫生组织的最新统计，从 2000 年起，全球道路交通死亡人数持续攀升，截至 2016 年，为 135 万，常年维持着 18 人/10 万人口的高死亡率形势。道路交通事故已成为 5～29 岁青年儿童的首要致死原因[1]。中德合作的《道路交通运输安全发展报告(2017)》中指出，2016 年我国共接报道路交通事故 864.3 万起，同比增加 65.9 万起，上升 16.5％。其中，涉及人员伤亡的道路交通事故 21 万多起，造成约 6.31 万人死亡；道路交通事故万车死亡率为 2.14，同比上升 2.9％[2]。同年，英、美、日的万车死亡率分别为 0.52、1.30、0.64[1]，与发达国家相比，我国交通安全水平仍有待改进。事故的特征研究和致因分析是提升交通安全的重要切入点，可为制定交通安全改善对策提供依据。

随着传感器功能的提升和车载数据记录仪的普及，事故的重现和致因推断不再只依赖于监控录像或当事人的自述，研究者们可以凭借自然驾驶数据，从更微观的角度(如驾驶行为)对事故进行深度分析。自然驾驶研究(naturalistic driving study, NDS)是指在自然状态下，利用高精度数据采集系统，观测、记录驾驶员真实驾驶过程的研究[3]。多源、实时、精确的自然驾驶数据能够为事故特征分析提供有力支持。但事故是小概率事件，需要通过长时间的观测才能得到足够的样本量。尤其在自然驾驶实验中，事故数不足以支撑个体驾驶员层面的统计分析。因此考虑用危险事件(safety-critical events, SCEs)作为事故替代指标。危险事件是任何需要驾

收稿日期：2019-04-13

基金项目：国家自然科学基金(51878498)；上海市科学技术委员会(18DZ1200200)

第一作者：王雪松(1977—)，男，教授，博士生导师，工学博士，主要研究方向为交通安全规划与政策、驾驶行为与车辆主动安全、驾驶模　　拟器应用、交通安全管理、交通设施安全分析。E-mail：wangxs@tongji.edu.cn

通信作者：徐晓妍(1995—)，女，硕士生，主要研究方向为交通安全、驾驶行为。E-mail：1831379@tongji.edu.cn

驶员做出避撞反应,且存在冲突对象和碰撞风险的情况,包括接近碰撞事件(near crashes)和碰撞事件(crashes,亦即事故)[4]。由于危险事件与事故的发生频率存在强相关性[5],且两者具有相似的因果机制[6],因此危险事件能够作为有效的事故替代指标,用于研究风险驾驶行为和推断事故致因。

自美国弗吉尼亚理工大学的100-Car和SHRP 2(Second Strategic Highway Research Program)自然驾驶研究项目开展以来,已有不少国外学者基于自然驾驶数据对危险事件进行了深入研究,包括探究危险事件的识别方法、分析危险事件的影响因素、利用危险事件进行驾驶员风险评估等。在危险事件识别方面,国外研究多采用传统的阈值法,即对车辆动力学参数设置阈值范围,从原始数据中自动识别符合条件的事件。这种方法的优势是保证了极少量的危险事件被漏报,但随之误报率大幅度提升,需要后期花费大量时间进行人工视频校对工作。

国内在建立危险事件的识别标准方面还存在较多空白。需指出的是,由于国内外驾驶环境不同,若直接照搬国外研究的阈值设定可能会导致识别效果不佳,因此亟需对国内的相关研究进行补充。上海自然驾驶研究(SH-NDS)由同济大学、通用汽车公司、弗吉尼亚理工大学三方合作,为国内首个自然驾驶研究项目。数据采集开始于2012年12月,结束于2015年12月,历时三年,共计19 133段出行,总行程161 055 km。该研究基于上海自然驾驶数据,建立危险事件的自动识别准则,从原始数据中提取可能的危险事件片段,在此基础上采用机器学习算法进一步过滤,在满足漏报率的同时,大幅度降低自动识别的误报率,从而减少后期人工校对的工作量。

1 研究综述

危险事件是任何需要驾驶员做出避撞反应的紧急情况,制动是最常见的避撞措施。Molinero等[7]基于欧洲5个国家的事故数据库,对不同场景的事故进行了深度分析。研究表明,60%的驾驶员在事故前会采取制动措施;Dingus等[4]利用100-Car自然驾驶数据,针对各种冲突类型的接近碰撞事件,统计了其中的避险措施类型。结果发现,超过80%的接近碰撞事件中,驾驶员通过及时踩下制动踏板成功避免了碰撞;紧急制动措施可用车辆纵向加速度的异常值(小于$-0.5g$)进行表征。除了纵向加速度,车辆速度、横向加速度、前向碰撞时间也常被用作识别危险事件的辅助依据。

目前大部分研究采用的危险事件识别过程如下:①对上述一系列车辆运动学参数(vehicle kinematics)设置阈值,从自然驾驶数据中自动提取可能的危险事件片段;②通过人工分析视频的方法,对初步识别得到的危险事件进行验证,筛选出有效的危险事件。既有研究中用于自动提取危险事件的车辆运动学参数如表1所示,满足任一类车辆运动学参数的阈值就会被识别为可能的危险事件。

表1 既有研究中危险事件提取准则

Tab.1 Summary of safety-critical event extraction criteria used in existing literature

既有研究	横向加速度/g	纵向加速度/g	紧急事件按钮	纵向加速度&前向碰撞时间		角速度/((°)·s^{-1})
				纵向加速度/g	前向碰撞时间/s	
Dingus等[4]	≥0.70	≥0.60*	触发	≤−0.50	≤4	3 s内振幅超过±4°
Lee等[8]	≥0.75	≥0.65*	触发	≤−0.65	≤4	3 s内振幅超过±4°
Hankey等[9]	≥0.75	≤−0.65	触发			0.75 s内振幅超过±8°
Perez等[10]	≥0.92	≤−0.75	触发且持续时间≥0.74 s			0.75 s内振幅超过±8°
Carney等[11]	≥0.55	≤−0.50				

注:*表示取纵向加速度的绝对值。

使用阈值法识别危险事件会导致较高的误报率,例如Dingus等[4]以及Perez等[10]识别危险事件的整体误报率均超过80%,需要在后期进行大量的人工校核和筛选工作。后续研究者提出了传统阈值法的改进算法。Sudweeks[12]在Dingus研究的基础上建立了一种角速度分类器,该分类器可过滤42%由角速度阈值识别到的无效事件。Wu等[13]提出了一种将人工校核视频工作量最小化的识别方法,使用阈值法初步筛选出可能的危险事件后,利用邹氏检验过滤掉与事故发生机理不同的事件;再通过生存分析和ROC(receiver operating characteristic)曲线确定车辆动态参数变化量的最佳阈值,进行第二轮自动筛选,最大幅度减少了留给人工校验的候选危险事件数。Kluger等[14]将离散傅里叶变换与k均值聚

类法结合，识别危险事件发生前后车辆加速度随时间变化的模式，运用该算法可将误报率降至22%。

也有研究者探索了阈值法以外识别危险事件的新方法。Dozza等[15]认为事件的危险程度应取决于驾驶员自身的感受和反应，利用多种图像处理算法对驾驶员面部视频进行分类，识别有效的危险事件。该方法可以覆盖84%的有效危险事件，各算法的平均误报率约为30%。Gao等[16]通过提取前向视频特征，生成每起事件的运动轮廓图（motion profile）；基于运动轮廓图和车辆动态学变量，建立多模态深度卷积神经网络用于识别危险事件。该方法可覆盖83%的有效危险事件，误报率控制在33%。

综上所述，目前国外学者用于危险事件识别的方法主要有以下三种：①传统阈值法；②结合分类算法改进传统阈值法；③图像识别算法。国内相关研究存在较多空白，亟需进行补充。既有研究都假设传统阈值法结合人工判别得到的危险事件是全样本，在传统方法基础上所作的改进都旨在降低误报率，减少人工判别的工作量，同时无法覆盖全样本，会产生一定的漏报率。因此本文认为，为了得到较为完整的危险事件集，阈值法不可舍弃；在传统方法基础上，需要寻求一种能同时降低误报率和控制漏报率的方法，过滤掉大部分无效事件。

支持向量机（support vector machine）模型是一种相对较新的机器学习模型，是Kecman[17]为了解决分类和回归问题而提出的。近年来，支持向量机模型被广泛应用于交通研究，包括交通流预测[18]、事件检测[19]、事故频率预测[20]等，具有较强的分类能力。因此本文考虑采用支持向量机在阈值法基础上对事件进一步分类。支持向量机模型的主要局限在于该模型像一个黑匣子，不能识别有效的解释变量。因此本文考虑利用随机森林模型筛选出重要特征，作为支持向量机模型的输入变量进行模型训练；并同时训练随机森林模型，与支持向量机模型的预测效果进行对比。

2　数据准备

本研究的数据来自“上海自然驾驶研究项目”，项目使用5辆配备了SHRP2 NextGen数据采集系统（包括4路摄影头、可跟踪前方8个物体的雷达系统、全球定位系统、车辆总线数据记录器等）的乘用车辆。数据采集系统的不同设备设置了不同的采样频率，分布在10～50 Hz[21]。数据采集系统在车辆点火后自动启动，熄火后自动关闭。数据采集开始于2012年12月，结束于2015年12月，历时3年，共计19 133段出行，总行程161 055 km。包括57位驾驶员，其中女性12位，男性45位。研究所用的驾驶员信息数据和车辆运行数据基本完整。

本文通过对车辆动态学参数（如横纵向加速度、前向碰撞时间等）设定阈值，从原始数据中提取可能的危险事件。初始阈值设置参考Dingus等[4]的研究。

（1）阈值类型1：横向加速度大于等于0.7g。

（2）阈值类型2：纵向加速度的绝对值大于等于0.6g。

（3）阈值类型3：紧急事件按钮触发。

（4）阈值类型4：横向加速度大于等于0.5g且前向碰撞时间小于等于4 s。

（5）阈值类型5：纵向加速度的绝对值大于等于0.5g且前向碰撞时间小于等于4 s。

只要某一时间戳的数据记录满足任一阈值类型，就会被自动识别为可能的危险事件，并提取该时刻前后10 s的视频记录用于人工校验。数据提取流程如图1所示。

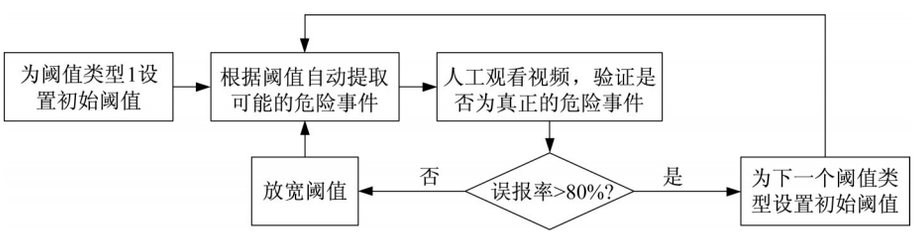

图1　危险事件提取流程

Fig.1　Safety-critical event extraction process

值得注意的是：采用较高的误报率（80%）是为了尽量减少遗漏的危险事件，确保充足的样本量。若满足以下任意两个条件，则人工判定为危险事件：①通过手部视频，发现驾驶员采取了紧急的避险操

作;②根据面部视频,发现驾驶员有明显的表情变化;③依据前向视频,发现自车与其他交通参与者或物体发生冲突。初始和最终阈值的设定如表2所示。

表2 事件提取阈值设定

Tab.2 Summary of extraction trigger criteria

阈值类型	初始阈值	最终阈值	样本增加量/%
1. 横向加速度	$\geqslant 0.7g$	$\geqslant 0.7g$	0
2. 纵向加速度	$\geqslant 0.6g$*	$\geqslant 0.5g$*	281
3. 紧急事件按钮	触发	触发	0
4. 横向加速度	横向加速度 $\geqslant 0.5g$;	横向加速度 $\geqslant 0.5g$;	0
& 前向碰撞时间	前向碰撞时间 $\leqslant 4$ s	前向碰撞时间 $\leqslant 4$ s	
5. 纵向加速度	纵向加速度 $\geqslant 0.5g$*;	纵向加速度 $\geqslant 0.45g$*;	207
& 前向碰撞时间	前向碰撞时间 $\leqslant 4$ s	前向碰撞时间 $\leqslant 4$ s	

注:*表示取纵向加速度的绝对值。

对于阈值类型1、3和4,设定为初始值时误报率已超过80%,因此不再进行调整。从表2可以看出,对于阈值类型2和5,通过放宽阈值,有效危险事件的样本量得到了大幅提升。利用阈值法共自动识别到3 623起可能的危险事件;人工校验后,将其中的591起认定为有效的危险事件,包括8起碰撞事件和583起接近碰撞事件。

3 方法与模型

利用阈值法识别危险事件仅能达到16.31%(591/3623)的准确率,增加了后期人工筛选的工作量。为改进识别方法,本文参照Wu等[13]"两轮筛选"的研究思路,考虑用阈值法进行初步过滤后,纳入机器学习方法进行深度筛选。基本流程如图2所示。首先对阈值法初步识别到的所有事件进行标签化处理(危险事件=1,一般事件=0),将事件标签作为输入变量;再将车辆动态参数统计量(如纵向加速度标准差)作为输入变量,分别采用随机森林模型和支持向量机模型识别危险事件。

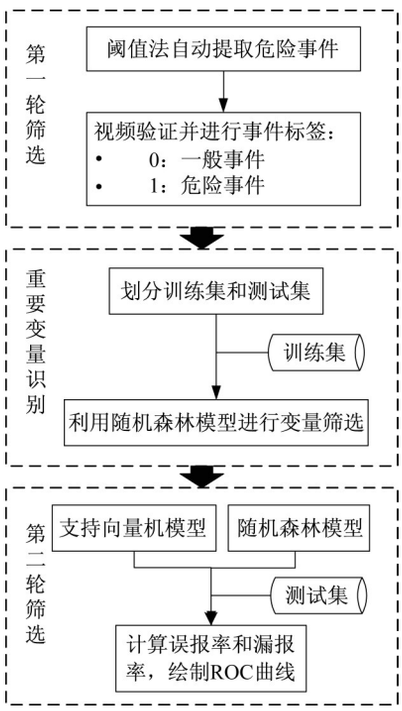

图2 运用机器学习的危险事件识别流程

Fig.2 Safety-critical event detection process using machine learning

3.1 机器学习输入变量

为确定有效的输入变量,首先需分析阈值法失效的原因。视频验证过程中三类常见的失效场景如下:①城市快速路或高速公路,由于路面颠簸或远处有车辆汇入主线,驾驶员在高速情况下本能地踩下制动踏板或转动方向盘,造成较大的横向或纵向加速度;②车辆接近交叉口时(无前车),本向绿灯转为红灯,为保证车辆不越过停车线,驾驶员采取紧急制动;③车辆经过下坡时,驾驶员为控制车速用力踩踏制动,导致某一时刻车辆的纵向加速度过大。

以上三类场景均不存在潜在的碰撞风险,但由于某一时刻的车辆运动学参数满足阈值条件,被错误地识别成危险事件。可见运动学参数的瞬时值不足以做出精确的判别。在选择机器学习的输入变量时,考虑纳入事件触发前后某一时域内,车辆动态参数(包括速度、横纵加速度、与前车的距离、与前车的速度差、前向碰撞时间)的统计值,包括最值、均值和标准差。输入变量汇总及计算时域如表3和图3所示。由于存在没有前车的情况,因此表3中的Δx、Δv和t_{TTC}三类变量可以为空值。图3为某一起事件在阈值触发前后共15 s内,各类运动学参数的时间序列图。对于该事件,运动学参数统计值的计算时域为纵向加速度最小值对应时刻t_0的前5 s和后3 s(图中阴影部分)。若事件由横向加速度阈值触发,

则t_0为横向加速度最大值对应的时刻。

表3　输入变量汇总

Tab.3　Summary of input variables

车辆动态参数	输入变量名
纵向加速度a_x(Xaccel)	Xaccel_min、Xaccel_max Xaccel_avg、Xaccel_std
横向加速度a_y(Yaccel)	Yaccel_min、Yaccel_max Yaccel_avg、Yaccel_std
自车车速v(V)	V_min、V_max V_avg、V_std
与前车的距离差Δx(Δx)	Δx _min、Δx _max Δx _avg、Δx _std
与前车的速度差[1)]Δv(Δv)	Δv _min、Δv _max Δv _avg、Δv _std
前向碰撞时间[2)] t_{TTC}(TTC)	TTC _min、TTC _max TTC _avg、TTC _std

1)为自车速度减去前车速度；2)为$\Delta x/\Delta v$。

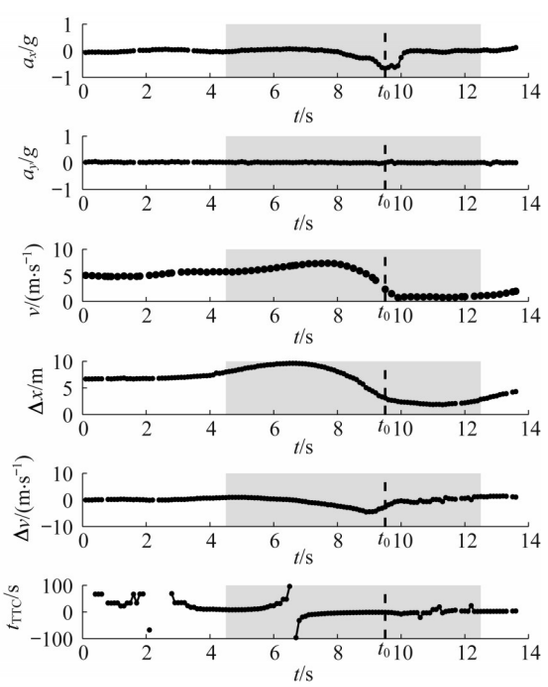

图3　输入变量的计算时域

Fig.3　Time horizon of input variable calculation

3.2　机器学习模型

3.2.1　随机森林模型

随机森林模型是由Breiman于2001年提出的一种机器学习算法[22]。其基本原理是：通过自助法(bootstrap)重采样技术，从大小为N的原始训练集中有放回地重复随机抽取N个样本，这N个样本组成一个训练样本集，一个训练样本集生成一棵决策树。决策树会从M个特征变量中随机抽取m个用于分裂节点。同样的过程重复k次，一个由k棵决策树组成的随机森林训练完毕。将测试集输入到每棵树中进行分类，最后由所有树对分类结果进行投票，投票数最多的即为最终分类结果。

由于每棵树是从大小为N的原始训练集中进行N次有放回采样，因此每棵树中会有重复的样本，同时也会有一些样本未被选中，这些未被选中的数据称为袋外数据B_{OOB}(out-of-bag，OOB)。若有k棵决策树，则随之会产生k个袋外数据。平均而言，每棵树进行放回抽样后，会有37%的数据没有被选中。推导公式如下：

当一棵树进行放回抽样后，某个样本一次也没有被选中的概率如下：

$$P(B_{\mathrm{OOB}})=\left(1-\frac{1}{N}\right)^N \tag{1}$$

当N趋近于无穷大时，$P(B_{\mathrm{OOB}})$会收敛到常量。证明如下：

$$\lim_{N\to\infty}P(B_{\mathrm{OOB}})=\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N=\frac{1}{\mathrm{e}}\approx 0.37 \tag{2}$$

随机森林模型不仅可以进行分类或回归，还能计算变量重要度M_{VIM}(variable importance measure，VIM)，帮助研究者筛选有效变量，降低数据维度[23]。M_{VIM}的计算是基于袋外数据分类准确率进行的。袋外数据分类准确率定义为：袋外数据自变量值发生轻微扰动后与扰动前的分类正确率的平均减少量。M_{VIM}的计算方式如下：

$$M_{\mathrm{VIM}}^{j}=\frac{1}{k}\sum_{t=1}^{k}\left(M_{\mathrm{P}tj}-M_{tj}\right) \tag{3}$$

式中：M_{VIM}^{j}表示第j个变量的重要度；k表示随机森林模型中的决策树数；M_{tj}和$M_{\mathrm{P}tj}$分别表示对第j个变量进行干扰前和干扰后，决策树t的袋外数据分类准确率。除了计算变量重要度，袋外数据还可用于选择每棵决策树分裂节点所需的最佳变量个数以及决策树数。

3.2.2　支持向量机模型

支持向量机模型的核心思想是：若一组二分类的数据有m个变量，则存在一个m维空间可以对这组数据进行表示。支持向量机模型的目标是在这个m维空间中寻找一个最能有效区分两类数据的$m-1$维超平面，即从众多超平面中寻找一个最优解。假设超平面服从线性方程，其表达式为

$$\left(x_1,x_2,\ldots,x_{\mathrm{m}}\right)\begin{pmatrix}w_1\\w_2\\w_3\end{pmatrix}+b=\boldsymbol{X}\boldsymbol{W}^{\mathrm{T}}+b=0 \tag{4}$$

式中：$\boldsymbol{X}$是输入变量组成的向量；$\boldsymbol{W}^{\mathrm{T}}$和$b$是待求的参数。根据推导[24]，SVM模型最终需解决以下最优

化问题：

$$\min_{W,b,\varepsilon}\left(\frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+C\sum_{i=1}^{N}\varepsilon_i\right) \quad (5)$$
$$\text{s.t.}\quad y_i\left(\boldsymbol{X}_i\boldsymbol{W}^{\mathrm{T}}+b\right)\geqslant 1-\varepsilon_i$$
$$i=1,2,\ldots,N$$
$$\varepsilon_i\geqslant 0$$

式中：ε_i为样本i的松弛变量，由于难以保证不同类型的数据点严格分布在超平面的两侧，松弛变量的引入放宽了约束条件，即使被错误地分在超平面的另一侧，只要样本点i至超平面的距离不超过ε_i，则仍满足约束条件；常数C为惩罚因子，由于ε_i越大，约束条件越弱，超平面的区分能力越弱，因此求取最优解的同时，也要使松弛变量之和尽量小，C决定了松弛变量之和的影响程度。

利用拉格朗日乘子法进行变换，式(5)变为

$$\max \boldsymbol{W}(\alpha)=-\frac{1}{2}\left(\alpha_i y_i\alpha_j y_j\boldsymbol{X}_i\times\boldsymbol{X}_j\right)+\sum_{i=1}^{N}\alpha_i \quad (6)$$
$$\text{s.t.}\quad 0\leqslant\alpha_i\leqslant C$$
$$\sum_{i=1}^{N}\alpha_i y_i=0$$

式中：α_i为拉格朗日乘子。

以上公式都是基于线性分类，即超平面服从线性方程。若线性分类无法解决问题，则需要进行非线性分类。其基本思想是：将原先的m维空间逐步映射到$m+1$维、$m+2$维、$m+3$维等更高维的空间，直到在某个更高维的空间中线性可分为止。所以，关键问题就变成了确定从低维坐标到高维坐标的映射关系。从式(7)中可以看出，样本点都是以两两内积的形式出现的，将样本点$\boldsymbol{X}_i$与$\boldsymbol{X}_j$的内积记作$k\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)$。因此上述的映射关系可以理解为样本点坐标在更高维度下的新的内积规则。这一规则就称为核函数。本文采用的核函数为高斯核(径向基函数)，其形式如下所示：

$$k\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\exp\left(-\frac{\left\|\boldsymbol{X}_i-\boldsymbol{X}_j\right\|}{2\sigma}\right) \quad (7)$$

式中：σ为核函数参数。综上可知，SVM模型共有两个待定参数(C,σ)。

3.3 模型预测效果

利用训练集训练得到随机森林和支持向量机模型后，对测试集进行预测，再基于分类准确率(classification accuracy，A_{cc})、误报率(false positive rate，R_{FP})、漏报率(false negative rate，R_{FN})以及受试者工作特征(receiver operating characteristic，ROC)曲线来对比两个模型的预测效果。本文所需处理的是一个二分类问题(是否为危险事件)，可能的分类结果如表4所示。

表4 二分类问题预测结果

Tab.4 Outcomes of a binary classification problem

实际结果	预测结果	
	1(危险事件)	0(一般事件)
1(危险事件)	真正类 (true positive，T_{P})	假负类 (false negative，F_{N})
0(一般事件)	假正类 (false positive，F_{P})	真负类 (true negative，T_{N})

依据表4，预测效果的度量指标计算如下：

(1) 分类准确率$A_{\mathrm{cc}}=(T_{\mathrm{P}}+T_{\mathrm{N}})/(T_{\mathrm{P}}+F_{\mathrm{P}}+F_{\mathrm{N}}+T_{\mathrm{N}})$。

(2) 误报率$R_{\mathrm{FP}}=F_{\mathrm{P}}/(F_{\mathrm{P}}+T_{\mathrm{N}})$。

(3) 漏报率$R_{\mathrm{FN}}=F_{\mathrm{N}}/(T_{\mathrm{P}}+F_{\mathrm{N}})$。

(4) ROC曲线的A_{uc}(area under the curve)值。

ROC曲线的横坐标为特异度(specificity)，取值为$1-R_{\mathrm{FN}}$；纵坐标为灵敏度(sensitivity)，取值为$1-R_{\mathrm{FP}}$。训练好的机器学习模型对每个测试样本都能得到一个预测概率。设阈值$p_0\in[0,1]$，若某样本的预测概率小于p_0，则归为一般事件；若大于p_0，则划分为危险事件。p_0取不同的值会产生不同的特异度和灵敏度，当p_0从0变化到1时，若干对特异度和灵敏度形成了ROC曲线。模型的预测效果可以由ROC曲线与坐标轴围成的面积A_{uc}进行度量。$A_{\mathrm{uc}}\in[0,1]$越大，说明预测效果越好。

4 结果与讨论

4.1 变量重要度排序

本文按照3:1的比例，将阈值法筛选出的3 623起事件随机划分成训练集和测试集。经过统计，在全样本、训练集和测试集中，危险事件的比例分别为16.31%、16.60%以及15.45%。为了避免数据集不平衡可能导致的误差，将训练集中的危险事件复制4份，尽可能保证危险事件与一般事件的比例为1:1。

利用随机森林模型进行变量重要度排序前，需要根据袋外数据误差确定随机森林模型中决策树的分裂节点特征变量数。从图4可以看出，当特征变量数目为5时，袋外数据误差达到最小，为0.033 2，因此可将结点特征变量数确定为5。

其次需要确定随机森林模型中的决策树数目。如图5所示，随着决策树数目递增，袋外数据误差逐

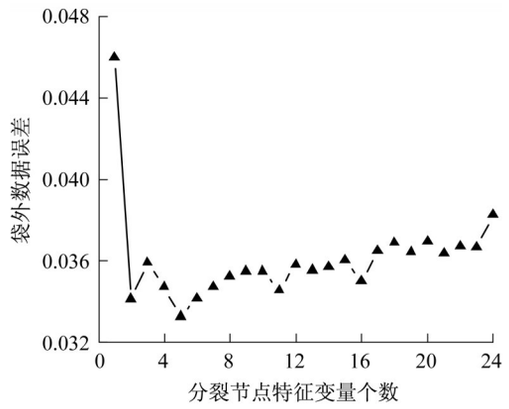

图4　分裂节点特征变量个数分析

Fig.4　Analysis of attributes number of split node

渐降低，并在650棵树后趋于稳定，因此将随机森林模型中决策树数量确定为650。

随机森林模型自身提供了两种变量选择方法：平均精确度减少（mean decrease accuracy）和平均节点不纯度减少（mean decrease in node impurity）。由于基于平均精确度减少的方法比基于节点不纯度减少的方法具有更好的非偏倚性能，因此既有文献中多采用前者进行变量筛选[25-27]。随机森林模型变量重要性排序如图6所示。从图6中可以看出，起到关键作用的变量有：纵向加速度的最小值、均值、标准差，与前车距离的最小值，车速的标准差，横向加速度的均值以及与前车速度差的均值。由于所有变量重要度的权重均大于1%，因此考虑将所有24个变量作为输入变量，放入机器学习模型中进行训练。

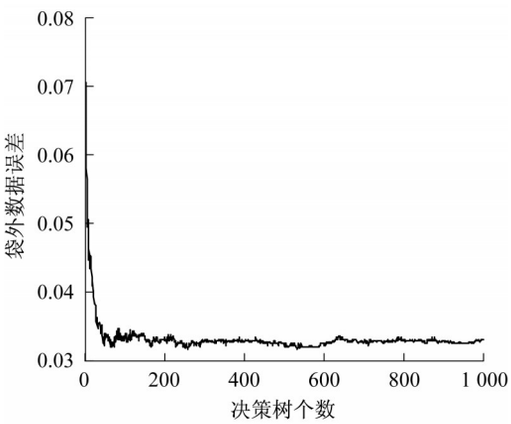

图5　决策树个数分析

Fig.5　Number analysis of decision tree

4.2　重要变量描述性统计

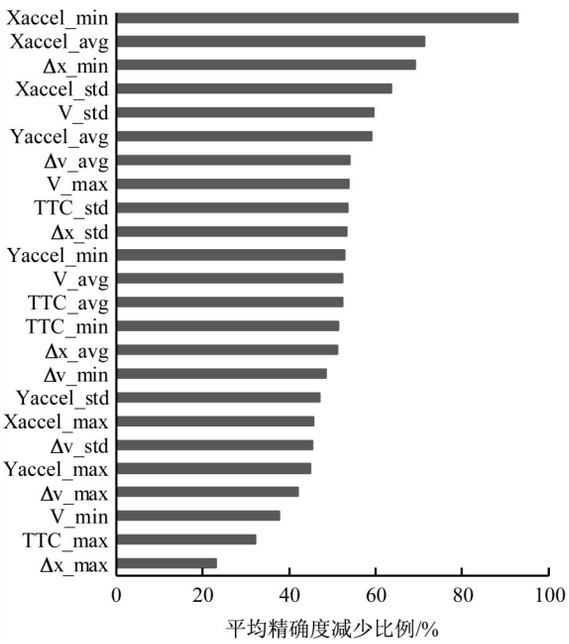

图6　变量重要度排序

Fig.6　Measurement of variable importance

表5　重要变量描述性统计

Tab.5　Descriptive statistics of important variables

变量名	训练集				测试集			
	一般事件		危险事件		一般事件		危险事件	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差
Xaccel_min	−0.42	0.49	−0.62	0.23	−0.44	0.52	−0.62	0.23
Xaccel_avg	−0.10	0.20	−0.11	0.12	−0.10	0.20	−0.10	0.11
Δx_min	12.58	20.17	4.56	6.11	12.94	19.21	4.04	5.67
Xaccel_std	0.09	0.07	0.16	0.05	0.09	0.07	0.16	0.05
V_std	3.72	3.48	3.62	2.22	3.84	3.53	3.58	2.30
Yaccel_avg	−0.01	0.05	−0.01	0.03	0	0.05	−0.01	0.03
Δv_avg	−1.56	3.19	−1.30	2.92	−1.40	3.30	−1.19	2.68
V_max	11.76	7.95	10.74	5.83	11.98	8.11	10.34	6.07
TTC_std	49.97	36.16	39.47	18.98	50.41	34.66	39.19	21.41
Δx_std	268.71	137.04	263.72	137.12	260.72	139.39	261.08	131.54

对重要度排序前10的变量进行描述性统计。表5汇总了训练集和测试集中，一般事件和危险事件的重要变量统计值。从表5中可以看出：①相比一般事件，危险事件发生期间的纵向加速度最小值

(Xaccel_min)更小,且标准差(Xaccel_std)更大,以上两个变量可以表征制动的紧急性;②危险事件发生期间,与前车距离的最小值(Δx_min)更小,速度差的均值(Δv_avg)更大。

4.3 模型结果

本文分别采用R语言中的"randomForest"以及"e1071"包来训练随机森林模型和支持向量机模型,基于测试集的分类准确率、误报率、漏报率以及A_{uc}值来评价预测效果。其中,随机森林模型的两个参数,即分裂节点特征变量个数以及决策树数个数已经在4.2节中明确,分别为5和650。根据3.2节,支持向量机模型有两个待定参数,惩罚因子C以及径向核函数参数σ,本研究采用R语言中的tune.svm函数进行十折交叉验证,对比训练集的分类误差,从而选取最佳的参数组合。结果表明,惩罚因子C取100,径向核函数参数σ取0.01时误差最小。

训练和预测后,两种机器学习模型的ROC曲线如图7所示。从图7中可以看出,支持向量机模型和随机森林模型的A_{uc}值都接近1,分别为0.897和0.896,说明两种模型均能达到较好的预测效果。

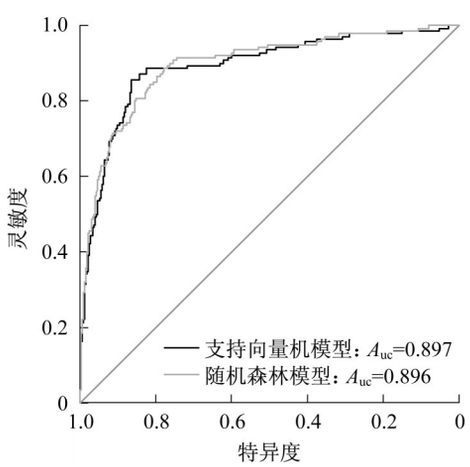

图7 两种机器学习模型的ROC曲线

Fig.7 ROC curves of two machine learning models

表6进一步展示了两种模型的预测结果。从表6中可以看出,随机森林模型和支持向量机模型的分类准确率均较高,分别为87.99%和86.09%。其中,随机森林模型的误报率较低,但漏报率很高,为37.14%,采用该算法容易损失较多的有效信息。支持向量机模型的误报率比随机森林模型高,却能将漏报率控制在12.86%,是一个可以接受的水平;且此时14.10%的误报率仍可以保证过滤超过85%的一般事件。因此针对本研究的目标,即尽可能降低自动识别的误报率,从而减少人工筛选的工作量,支持向量机模型的预测结果更优。

表6 两种机器学习模型的预测效果对比

Tab.6 Comparison of performance of two machine learning models

模型	预测效果度量指标			
	A_{cc}/%	R_{FP}/%	R_{FP}/%	A_{uc}
随机森林	87.99	5.22	37.14	0.896
支持向量机	86.09	14.10	12.86	0.897

对比本文的支持向量机模型与既有文献中的阈值法改进算法,结果如表7所示。需指出的是,进行对比的3篇文献采用的数据来源均为自然驾驶数据,与本文的数据结构一致;且数据采集频率以及阈值法提取危险事件采用的车辆运动学特征也相似,因此认为具有一定的可比性。从表7中可以看出,本研究使用的支持向量机方法在误报率和漏报率方面都优于其他研究的预测结果。

表7 支持向量机模型与其他模型的预测效果对比

Tab.7 Comparison of prediction performance of SVM and models in literature

作者	方法	R_{FP}/%	R_{FP}/%
本文	支持向量机	14	13
Sudweeks等[12]	角速度分类器	58	17
Dozza等[15]	面部视频识别	30	16

5 结语

基于上海自然驾驶数据,依据横纵向加速度和前向碰撞时间的瞬时值,建立危险事件的自动提取阈值标准,从原始数据中识别出3 623起可能的危险事件。经人工验证,其中591起为有效的危险事件。为降低阈值法过高的误报率,减轻后期人工校对的工作量,采用机器学习对阈值法初步识别的事件进行深度筛选,主要步骤如下:①按照3:1的比例,将3 623起事件随机划分为训练集和测试集。②基于训练集,利用随机森林模型识别重要的车辆动态参数特征,将其作为输入变量训练随机森林模型和支持向量机模型。③对测试集进行预测,计算误漏报率。

结果表明:①起到关键作用的变量有纵向加速度的最小值和均值、与前车距离的最小值以及车速的标准差。②相比随机森林模型,支持向量机模型的预测效果更优,在控制漏报率的同时,可过滤85.9%的无效事件。研究采用的方法可大幅度提升

危险事件的识别效率，可为基于自然驾驶数据识别危险事件的后续研究提供一定参考。

参考文献：

[1] World Health Organization. Global status report on road safety 2018[M]. Geneva：World Health Organization，2018.

[2] GSUL M，胡予红，周旋，等 . 道路交通运输安全发展报告(2017)[J]. 中国应急管理，2018(2)：48.
GSUL M，HU Yuhong，ZHOU Xuan. Road traffic safety development report (2017) [J]. China Emergency Management，2018(2)：48.

[3] FITCH G M ，HANOWSKI R J. Using naturalistic driving research to design，test and evaluate driver assistance systems [M]// Handbook of Intelligent Vehicles. London：Springer，2012.

[4] DINGUS T A，KLAUER S G，NEALE V L，*et al*. The 100-Car naturalistic driving study. Phase 2：results of the 100-Car field experiment [R]. Washington D C：Department of Transportation. National Highway Traffic Safety Administration，2006.

[5] WU K F，AGUERO-VALVERDE J，JOVANIS P P. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level [J]. Accident Analysis & Prevention，2014，72：210.

[6] GUO F，KLAUER S G，HANKEY J M，*et al*. Near crashes as crash surrogate for naturalistic driving studies [J]. Transportation Research Record，2010，2147(1)：66.

[7] MOLINERO A，EVDORIDES H，NAING C L，*et al*. Accident causation and pre-accidental driving situations. Part 2. In-depth accident causation analysis. Deliverable D2.2 [R]. Leicester：Loughborough University，2008.

[8] LEE S E，SIMONS-MORTON B G，KLAUER S E，*et al*. Naturalistic assessment of novice teenage crash experience [J]. Accident Analysis & Prevention，2011，43(4)：1472.

[9] HANKEY J M，PEREZ M A，MCCLAFFERTY J A. Description of the SHRP 2 naturalistic database and the crash，near-crash，and baseline data sets [R]. Blacksburg：Virginia Tech Transportation Institute，2016.

[10] PEREZ M A，SUDWEEKS J D，SEARS E，*et al*. Performance of basic kinematic thresholds in the identification of crash and near-crash events within naturalistic driving data [J]. Accident Analysis & Prevention，2017，103：10.

[11] CARNEY C，MCGEHEE D V，LEE J D，*et al*. Using an event-triggered video intervention system to expand the supervised learning of newly licensed adolescent drivers [J]. American Journal of Public Health，2010，100(6)：1101.

[12] SUDWEEKS J D. Using functional classification to enhance naturalistic driving data crash/near crash algorithms [R]. Blacksburg：Virginia Tech Transportation Institute，2015.

[13] WU K F，JOVANIS P. Screening naturalistic driving study data for safety-critical events [J]. Transportation Research Record：Journal of the Transportation Research Board，2013 (2386)：137.

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

[15] DOZZA M，GONZÁLEZ N P. Recognising safety critical events：can automatic video processing improve naturalistic data analyses? [J]. Accident Analysis & Prevention，2013，60：298.

[16] GAO Z，LIU Y，ZHENG J Y，*et al*. Predicting hazardous driving events using multi-modal deep learning based on video motion profile and kinematics data[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Hawaii：IEEE，2018：3352-3357.

[17] KECMAN V. Support vector machines - an introduction[M]// Support Vector Machines：Theory and Applications. Berlin，Heidelberg：Springer，2005.

[18] ZHANG Y，XIE Y. Forecasting of short-term freeway volume with v-support vector machines [J]. Transportation Research Record：Journal of the Transportation Research Board，2008 (2024)：92.

[19] CHEN S，WANG W，VAN ZUYLEN H . Construct support vector machine ensemble to detect traffic incident [J]. Expert Systems with Applications，2009，36(8)：10976.

[20] LI X，LORD D，ZHANG Y，*et al*. Predicting motor vehicle crashes using support vector machine models [J]. Accident Analysis & Prevention，2008，40(4)：1611.

[21] 王雪松，杨敏明 . 基于自然驾驶数据的变道切入行为分析 [J]. 同济大学学报 (自然科学版)，2018，46(8)：1057.
WANG Xuesong，YANG Minming. Cut-in behavior analysis based on naturalistic driving data [J]. Journal of Tongji University (Natural Science)，2018，46(8)：1057.

[22] BREIMAN L. Random forests [J]. Machine Learning，2001，45(1)：5.

[23] VERIKAS A，GELZINIS A，BACAUSKIENE M. Mining data with random forests：a survey and results of new tests[J]. Pattern Recognition，2011，44(2)：330.

[24] CORTES C，VAPNIK V. Support-vector networks [J]. Machine Learning，1995，20(3)：273.

[25] STROBL C，BOULESTEIX A L，ZEILEIS A，*et al*. Bias in random forest variable importance measures：illustrations，sources and a solution [J]. BMC Bioinformatics，2007，8 (1)：25.

[26] CALLE M L，URREA V. Letter to the editor：stability of random forest importance measures [J]. Briefings in bioinformatics，2010，12(1)：86.

[27] NICODEMUS K K. Letter to the editor：on the stability and ranking of predictors from random forest variable importance measures[J]. Briefings in Bioinformatics，2011，12(4)：369.