文章编号:1005-0523(2017)02-0029-08

基于决策树方法的偏远地区山区公路驾驶安全性研究

李 卓,陈雨人

(同济大学交通运输工程学院,上海 201804)

摘要:山区公路复杂的组合驾驶环境因素与事故数据的缺乏记录都使得分析山区公路驾驶安全性十分困难。建立了一套完整的山区公路驾驶安全性实验方法,并提供了一种结合了随机森林法则与基于机器学习决策树的方法,以分析山区公路上导致驾驶事故的主要因素。实验团队使用行车记录仪结合视频识别技术以获取沿途的道路驾驶环境与驾驶员在试验中的驾驶行为。用"最大信息熵增长率"法结合AIC准则以分析归纳模型中包含的主要驾驶环境因素。结论显示:急窄弯道、穿行于村镇间、缺乏视距是影响驾驶员在山区公路上安全行驶的主要原因。证明了这种方法可以用于指导山区公路的设计,提高山区公路上的行驶安全性。

关键词:驾驶安全性实验;驾驶环境因素;决策树分类

中图分类号:U411 文献标志码:A

DOI:10.16749/j.cnki.jecjtu.2017.02.004

山区公路在运输、旅游和村镇往来路线上往往扮演着重要的角色。但是近年来,由于行车环境的不协调,山区公路上出现了越来越多的交通事故与潜在的危险路段。现阶段,山区公路的设计理论趋于规范化[1],且被普遍应用于各类山区公路设计中。但是山区公路往往具有复杂的环境条件,仅用规范化设计理论难以避免一些路段事故频发。因此,许多学者纷纷开展了相关的研究,如:车速限制研究[2-4],结合交通量,道路环境,事故数据的交通事故成因分析,统计建模以进行事故预测等[5]。以上技术路线可以总结为归纳演绎法。在归纳学习法中,研究者通过观测数据样本,进行其他场景下的预测,以得到普适性的结论。许多不同的方法应用于分析驾驶安全性,综述研究[6]中已经做了详尽的总结:Probit模型,Logit模型,Log-linear模型等等。数据挖掘技术是一种分析大量数据并且将之转换有用信息与知识的方法。人工神经网络与一些分类、回归方法已经被普遍地应用于道路安全分析中。但是偏远地区的山区公路经常远离目前的道路信息系统监管范围,而且相关的事故数据也难以获得。目前在以往的研究中,针对这类山区公路上的驾驶安全性分析比较少。因此,在分析驾驶行为安全性之前,必须通过实地调查与实验获取研究道路上的驾驶环境变量。

因此本研究的目的是提出一种针对缺乏事故数据的山区道路的研究方法,并且将研究成果应用于消除山区公路上的事故多发段。作为案例,本研究中的研究方法已经应用于国道318,并且根据研究结论已经完成了318国道安全性评估。余下篇幅将从以下内容展开:首先罗列研究框架,然后介绍研究中所用的数据集,详细介绍实验路段的特点与实验方法,最后对于研究做结论总结。本研究提出了一种决策树方法用于分析道路驾驶环境因素对驾驶者的影响机理。

1 研究框架

目前已有许多对车流和道路几何特性相关的变量与交通事故的评估, 以及交通事故预测的试验方法。如果事故数据十分详尽,包含了各种属性变量,那么可以使用聚类分析以研究不同类型的交通事故。但是这种方法不能用于本研究的道路。因此,首先需要采集观测数据,再进行数据分析。

收稿日期:2016-09-03

基金项目:国家科技支撑计划课题(2014BAG01B06)

作者简介:李卓(1992—),女,硕士研究生,主要研究方向为道路安全与环境。

指导老师:陈雨人(1966—),男,博导,研究方向为道路规划与设计、交通安全与环境等。

决策树是数据挖掘技术中的一种[7]。决策树方法中包含了两种数据集(属性变量与分类变量)。需要根据已有研究标准定义不同路段的道路驾驶安全等级,然后将安全等级作为分类变量进行数据挖掘。此外,属性变量应该从试验中获取。所以本研究的框架为:

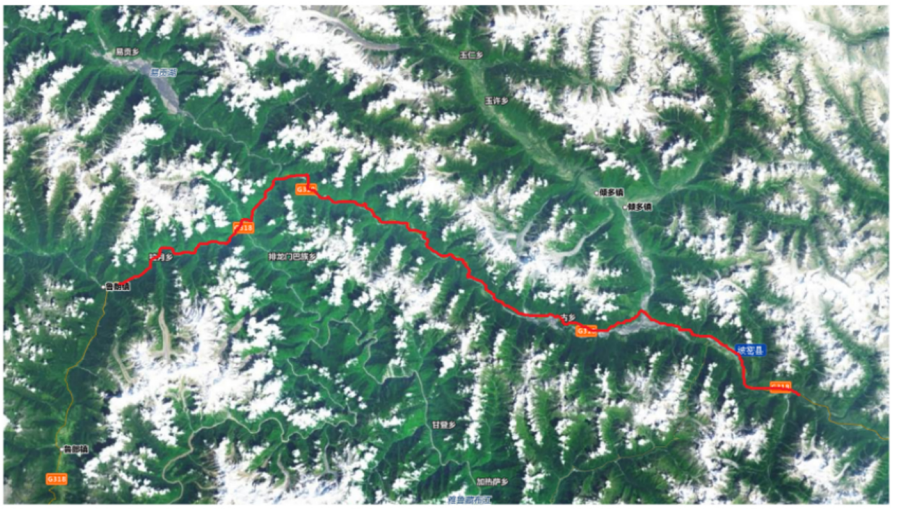
1) 实地实验,记录实验数据;

2) 对实验过程进行安全性评估(分类变量),提取驾驶环境因素(属性变量);

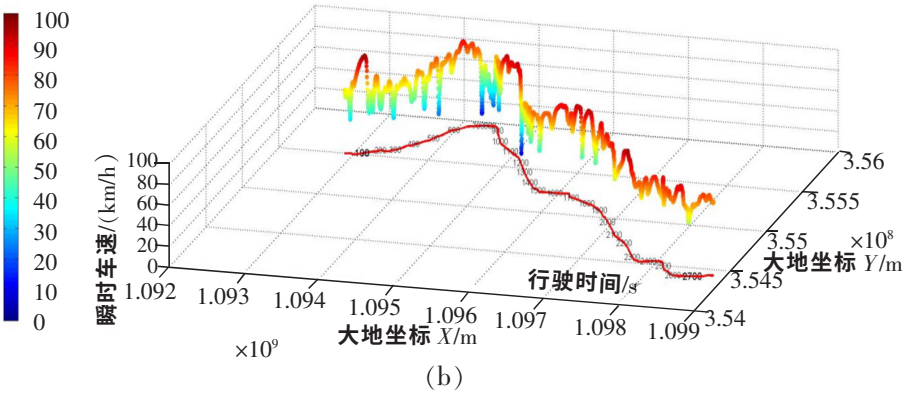
3) 决策树方法分析并总结结论。

2 数据集

国道318位于偏远的高海拔地区。全实验路段的总里程为52 569 m。道路为双向两车道,但是中央分割线不全,部分路段双向行驶车道分割模糊。道路地图如图1(a)所示。可见,道路沿线上经过了一些村庄与山谷线路段。道路路况艰难,周边环境复杂。根据调查,该试验道路上有多处事故高发路段,但是具体的统计信息与事故段位置无从获知。因此,研究团队远赴该地区,进行了实地实验以采集道路行驶环境因素,以期对于该路段的驾驶安全性进行评估。实验与数据处理过程将在下文介绍。



(a)



(b)

图1 国道318灵芝地区段与该路段上的实验车速

Fig.1 National road 318 in Linzhi areas and the travel speed recorded

2.1 实验目的与实验方法

为了全面地观测该试验道路,本研究实施了一系列实车行驶实验。实验过程简述如下:在良好天气条件下,让驾驶员沿路进行自然驾驶,使用一辆装载了行车记录器(GARMIN GDR35)的测试车。行车记录器与全球定位系统(GPS)相结合,并且完美地将GPS信号与摄像机镜头结合。这使得它可以同时记录车辆位置、车

速、加减速度与驾驶员的视野内容。采集的信息将以 1 Hz 的频率记录，采集的视频分辨率为 1 920*1 080 像素，帧率 30 帧/s，摄像机镜头焦距为 f=2.0，视野角度为水平 51.222 25°，垂直 34.998 939°。

道路环境复杂如图 1(b)所示，即便实验中的驾驶员已经有十几年的户外驾驶经验与山地驾驶技巧，但是他仍然难以保持稳定的驾驶车速。让驾驶员行驶在道路的右侧车道上，但是就行车记录仪所记录的结果来看有时这很难做到。没有指定实验期望车速，而且该道路交通量极小，因此可以认为实验符合自然驾驶的标准。

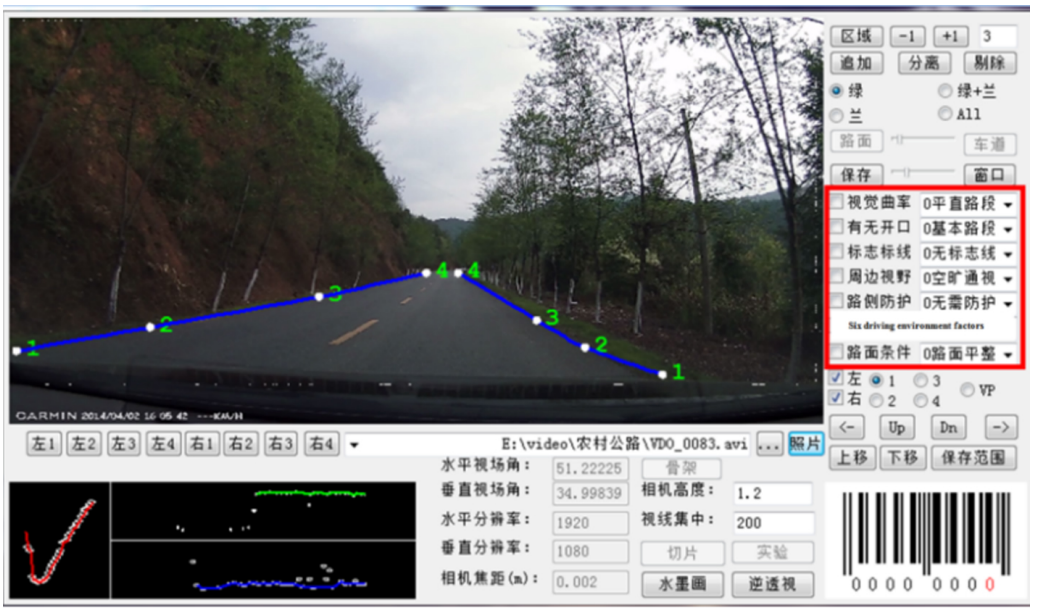


图 2　驾驶视频识别软件

Fig.2　Driving video recognition software

2.2　实验数据处理与驾驶安全性评估

如图 2 所示，本研究使用了一款自主开发的视频识别软件以识别驾驶全感知前方道路上的驾驶环境信息。该视频识别技术基于现有研究中的模型。该视频识别软件可以提取出 6 种驾驶环境因素(是否是急陡弯(IFSC)(评判标准是平均圆曲线半径小于 200 m)，是否经过村镇(IFTV))，是否有中央分隔线(IFCL)，是否有足够的时距(IFSD)，是否有安全护栏(IFRR)与道路是否颠簸(IFBR))，基于驾驶员的视觉车道模型 Cat-Mull-Rom spline。这款软件还可以识别圆曲线半径，坡道坡度，行驶里程等。每秒驾驶环境因素被识别为 6 个哑元变量(0 表示未发生，1 表示发生)。由于瞬时记录的驾驶环境因素需要经过驾驶员感知，思考，反应到驾驶行为中去，所以数据分析的时间尺度需要扩大。将试验路段根据直缓点，缓圆点，圆弧中点，圆缓点，缓直点分为 85 个路段。每个路段平均长 656 m，平均行程时间 35 s。在每段路段上，计算 6 个哑元变量的距离占有率，以反应 6 种影响因素的强弱。结合每段路段的平均圆曲线半径(ACR)与平均坡度(ASR)，对于每段路段都有 8 个变量。

根据上文的数据处理方法，需要根据 8 个变量评估每段路段的驾驶安全性。本研究使用路段上的车速记录来评估各个路段的驾驶安全等级作为分类变量。通过行驶车速可以评估交通事故潜在风险。在已有的研究中，相邻路段的车速差是最关键的评估因素(ATS)。良好的设计与行驶环境可以使得驾驶员以平稳、连续、安全的车速行驶，从而降低了发生交通事故的可能性。目前已有几种关于 $\triangle v$ 的评估标准。参考 V85%标准[2]，以相邻路段的平均车速差作为评估标准。该标准基于驾驶员驶过相邻两个路段的平均车速($\triangle_a^i$ 对于第 i 个路段)的绝对差值。该标准如表 1 所示。

表1 ATS标准
Tab.1 ATS criterion

$\triangle v=\lvert v^{i}_{a}-v^{i-1}_{a}\rvert$(km/h)	安全等级
$\triangle v\leqslant 10$	无风险
$10\leqslant\triangle v\leqslant 20$	中等危险
$\triangle v\geqslant 20$	十分危险

另一种建立驾驶行为与驾驶安全性的标准为车速下降系数法SRC[2]。

$$SRC=\frac{v^{i}_{a}}{v^{i-1}_{a}} \tag{1}$$

这种方法考虑v_i(当车辆从第i个路段驶出,进入第i+1个路段时的瞬时车速)与a_i(车辆驶出第i个路段驶出,进入第i+1个路段时的瞬时减速度)。这种标准如表2所示。

表2 SRC标准
Tab.2 SRC criterion

准则危险等级	v_i<60 km/h指标	60 km/h<v_i<80 km/h指标
无危险	$0.8<SRC \parallel a_i<1.5$	$0.9<SRC \parallel a_i<1.5$
中等危险	$0.8\geqslant SRC\geqslant 0.6 \parallel 2.5\geqslant a_i\geqslant 1.5$	$0.9\geqslant SRC\geqslant 0.7 \parallel 2.5\geqslant a_i\geqslant 1.5$
十分危险	$SRC<0.6 \parallel 2.5<a_i$	$SRC<0.7 \parallel 2.5<a_i$

同时使用ATS标准与SRC标准对实验数据的驾驶危险性进行了分类。

3 决策树分类方法

3.1 问题描述与假设

为了方便本研究反应问题,提出了如下基本假设:

1)为了分析许多驾驶环境因素的影响,这个问题转化为多等级分类问题,定义1—3对应3种不同的驾驶安全性等级。

2)所有的训练集假设由独立同分布的样本构成,因此他们会单独进入决策树的构建。

定义本研究中的变量如表3。

表3 变量设计
Tab.3 Variable design

x_i,y_i	第i个训练集中的样本,第i个分类集中的样本
a_j	一个样本的第j个属性变量
$TreeGenerate(D,A)$	建立一个节点的函数(输入为训练集D,属性集A)
a_o,a^s_o	分割中使用的最优属性变量,与这个变量的第s个值
D_s	有与属性a_o中的a^s_o相同值的D的子集
p_k	当前训练集D中归为第k类的样本的比例
$Node$	当前训练集D中的结点
$Ent(D)$	D的信息熵,以表征D的纯度
$Gain(D,a)$	以属性a进行分割时D的信息熵增量
$IV(a)$	属性a的内在值
$Gain\_ratio$	信息熵增加率
$Numclass_i,Numclass_{vcted}$	森林中投票给第i类的票数,森林中投票胜出的决策树个数
N_{class}	分类总数量

决策树采用分而治之的策略，决策树过程为递归过程，树上每个节点不但可以评估驾驶的危险性，还可以找出导致不安全驾驶行为的关键因素。因此决策树方法适用于分析多驾驶环境因素的组合影响。决策树的末端分支将通过这些驾驶环境因素指向不同的驾驶安全性分类。

3.2　决策树算法流程

使用经典的 C4.5 方法[10]训练观测样本。算法流程如下：

1）数据训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 与属性变量集 $A=\{a_1,a_2,\cdots,a_d\}$；

2）$TreeGenerate(D,A)$ 创建树节点记为 Node；

3）If 如果 D 中的样本已经全被分为类 C then

　　标记 Node 为 C 的叶节点；return

end if

4）If A 中所有的属性变量已经被剔除 OR 所有的训练集 D 中的样本在 A 属性值上取值相同 then

　　将 Node 标记为叶节点，并且将其类别标记为 D 中样本数最多的类 return

end if

5）选择最优的划分属性变量 a_o，随后在属性集中剔除该属性 a_o；

$$a_o=\arg\max_{a\in A} Gain\_ratio(D,a) \tag{2}$$

6）for a_o 在当前 D 中的每一个值 do

为 Node 创建一个分支并选出 D_s；

if D_s 为空 then

将 Node 标记为叶节点，将它的类别标记为 D 中样本最多的类 return；

else

$TreeGenerate(D,A\backslash\{a_o\})$ 创建分结点；

　end if

end for

7）输出一个以 Node 为根节点的决策树

在第 5 步中，使用“最大信息熵增加率”规则选取最佳分类属性划分样本集。这个过程解释如下。

1）以样本集信息熵表征其纯度。

$$Ent(D)=-\sum_{k=1}^{|\gamma|}p_k\log_2 p_k \tag{3}$$

信息熵越小，样本集纯度越高。因此，使用信息熵增加率较高的属性划分样本集，可以提高样本集纯度，更好的界定危险驾驶行为的分级机理。

2）使用信息熵增加率表征样本集纯度的增加。

$$Gain(D,a)=Ent(D)-\sum_{f=1}^{F}\frac{|D_f|}{|D|}Een(D_f) \tag{4}$$

$$IV(a)=-\sum_{f=1}^{F}\frac{|D_f|}{|D|}\log_2\frac{|D_f|}{|D|} \tag{5}$$

$$Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)} \tag{6}$$

当前样本集中仅剩一个样本时，划分终止。

3）使用文献[10]中提出的一种启发式算法，寻找一系列须有高信息熵增加率的 $a_o s$，然后选出当前具有最大信息熵增加率的属性 a_o。

为使分类器的结果偏差尽可能小，将 C4.5 算法与随机森林法则结合。每次决策树训练，随机选取样本

集中的85%样本,进行训练并得出一颗决策树。如此进行10次,得出10棵决策树。在测试数据的分类过程中,10棵树得出10组不同的结果,通过众数投票得出每一个样本的分类结果。对投票过程记性了统计,定义投票一致性(VV)以表征模型的稳定性。VV越小表明10决策树倾向于得出不同的结论一致性越差,这也说明了模型的分类效果较为不稳定。

$$VV=\frac{\sqrt{\sum_{i=1}^{Nclass}(Numclass_i-Numclass_{voted})^2}}{10\times N_{class}} \tag{7}$$

3.3 结果验证方法

为了验证模型的稳定性与准确性,使用k-fold交叉验证法。总体样本集被划分为k份。选择其中一份作为测试数据,其余$k-1$份作为训练数据。如此交叉进行k次测试,以保证在全样本范围内对模型进行有效检验。本研究中k取5,最终以5次测试的准确性验证该模型。总体算法流程图如图3。

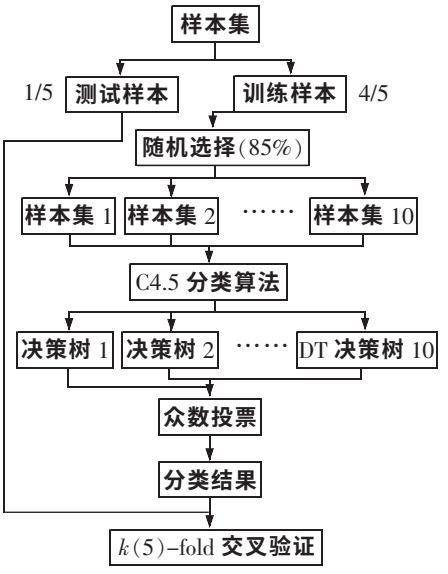


图3 算法流程图

Fig.3 Algorithm flow chart

3.4 变量影响作用度量

纳入该模型的变量重要性度量如下。

$$VIM(a)=\sum_{i=1}^{h}\frac{na_i}{N_{class}}Gain\_ratio(D,a=a_i) \tag{8}$$

其中,C为分类种类,本研究共3类;na_i为属性值数量($a=a_i$)。$VIM(a)$表征了当使用属性值a时,在类别C上的信息熵增量。VIM的值表征了每个变量标准化的重要性度量,通过该指标可以得出纳入模型的各个属性之间的优先度排序。

4 结果分析

基于MATLAB软件,以上算法得以实现。为了使得决策树中的划分易于理解与解释,本研究限定决策树最高4层。

4.1 变量筛选

根据最大信息熵增加率规则,最初被选用于划分的属性成为根节点,其对于诊断驾驶安全性分类最为有效。根据表4统计,路段上的急陡弯、视距不足、经过村镇是3类最主要的影响驾驶安全性的道路环境因素。

表4 每种属性变量成为根节点的频数统计

Tab.4 The frequency of each attribute variable developing into root-node

指标	IFSC	IFSD	IFTV	ACR	IFBR	IFCL	ASR	IFRR
频数	42	24	20	17	16	10	6	5
序号	1	2	3	4	5	6	7	8

变量筛选的目的为找出可以最恰当解释与训练样本集的属性变量。此外,通过对于8个属性变量的VIM值计算,可以根据变量的标准化重要度得出进入模型的优先度排序。确定排序后,本研究通过Akaike

Information Criterion，AIC[11]法则确定进入模型的最佳变量数量。AIC 准则是为了权衡模型复杂度与模型精度，避免模型发生过拟合现象，其 AIC 指标计算方法如下。

$$AIC=2K+n\ln(RSS/n) \tag{9}$$

式中：RSS 是模型测试中的均方差；K 是进入模型的变量数；n 为样本集观察数。

由图 4 可见，当变量数为 3 时，AIC 值最小，此时模型的复杂度最为适宜。根据 VIM 值确定出进入模型的 3 个属性变量为 IFSC, IFSD，IFTV。这个结果与根据最大信息熵增加率规则确定的 3个最主要影响因素的结果吻合。

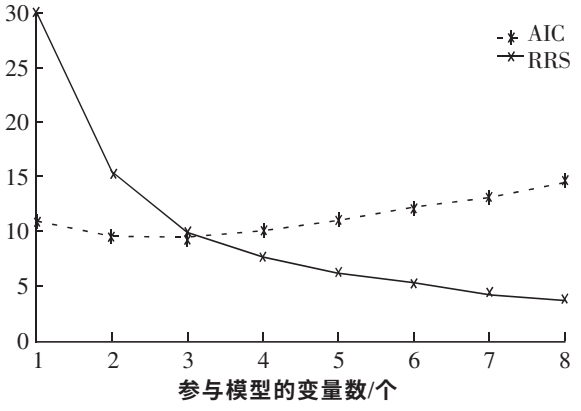


图 4　AIC 值与 RRS 值

Fig.4　AIC values and RRS values

4.2　模型训练与测试结果

在 SRC 标准下：8 参数模型的 5 组测试平均准确性与 VV 值分别达到了 0.86 与 0.73。如果仅使用以上 3 参数模型，平均准确性与 VV 值分别达 0.80 与 0.71。同理在 V85%标准下：8 参数模型的 5 组测试平均准确性与 VV 值分别达到了 0.89 与 0.75。3 参数模型平均准确性与 VV 值分别达 0.78 与 0.69。说明 3 参数模型已可以对多数驾驶安全性的变化情况作出解释，精度较高，模型内部统一性较高。

5　结论与总结

通过实地测试对于偏远地区山区公路驾驶安全性做了分析。使用了行车记录仪记录驾驶员对环境的视觉信息，并通过视频识别技术从中提取出 8 种驾驶环境影响因素。同时根据实验车辆行驶轨迹对路段行驶安全性进行分级。建立了决策树算法分析了组合驾驶环境因素对于山区公路监视安全性的影响。

从标准化变量重要度和参与划分的统计频数两个方面均证实了小半径曲线（<200 m）、停车视距不足与车辆穿过村庄这 3 个因素对于驾驶安全性影响最大，进一步研究了组合因素（诸如：安全栏缺失，中心线确实，等）的影响，并训练得出了较为精确的决策树模型，解释了驾驶环境因素对路段驾驶安全性的影响机理。本研究中的试验与分析方法不需要事故数据，通过该实验路段所训练得出的决策树可以用于其他山区公路的驾驶安全性分析，并得出各个路段的安全性分析结果，指导山区公路的设计与设施改善。

参考文献：

[1] 杨志清，郭忠印，杜晓丽. 基于视觉信息的高速公路运行车速预测模型[J]. 同济大学学报：自然科学版，2007，35(7)：929-934.

[2] 马社强，刘东，路峰. 车速对交通安全的影响及管理研究[J]. 公路交通技术，2008(5)：139-142.

[3] 陈涛，魏朗. 道路行车安全性虚拟评价方法研究[J]. 安全与环境学报，2006，6(6)：115-118.

[4] 孟妮，韩丹. 聚类分析和模糊逻辑在驾驶行为辨识中的应用[J]. 计算机与数字工程，2013，41(7)：1097-1099.

[5] WANG Y，SHEN D，TEOH E K. Lane detection using spline model[J]. Pattern Recognition Letters，2000，21(8)：677-689.

[6] SAVOLAINEN P T，MANNERING F L，LORD D，et al. The statistical analysis of highway crash-injury severities：A review and assessment of methodological alternatives[J]. Accident Analysis & Prevention，2011，43(5)：1666-1676.

[7] HAN J，KAMBER M，KAMBER M. Data Mining: Concepts and techniques，morgan kaufmann. Machine Press，2001 in Chinese，2006，5(4)：394-395.

[8] 王岳,陈雨人. 低等级公路交叉口与城市道路非信号灯交叉口车辆运行速度特征比较分析[J]. 华东交通大学学报,2016,33(3):79-86.
[9] 陈雨人,余博,贺思虹. 基于视觉感知偏差的公路几何平纵协调性分析技术[J]. 同济大学学报:自然科学版,2015,43(9):1347-1354.
[10] QUINLAN J R. C4.5:programs for machine learning[M]. Morgan Kaufmann Publishers Inc,2014:62-79.
[11] AKAIKE H. Factor analysis and AIC[J]. Psychometrika,1987,52(3):317-332.

Analysis of Driving Safety on Mountain Highway in Remote Areas Based on Decision Tree Method

Li Zhuo, Chen Yuren

(College of Transportation Engineering, Tongji University, Shanghai 201804, China)

Abstract: The combination of complex driving environment factors and the lack of accident data make the analysis of driving safety on mountain highway difficult. This paper proposed a complete set of experimental methods for driving safety on mountain highway and provided a decision tree method combining with Random Forest method to identify the main factors leading to accidents. By using automobile data recorder and radio frequency identification, the experimental team obtained information about the driving environment and drivers' behaviors. The approach of maximum gain ratio and Akaike information criterion are adopted to analyze and summarize the main driving environment factors involved in the model. The results showed that sharp road curvatures, traveling through villages and lack of sight distance are the main factors for driving safety on mountain highway. It is confirmed that this approach can provide reference for the design of mountain highway to improve driving safety.

Key words: driving safety experiment; driving environment factors; decision tree method

(责任编辑　王建华)