

文章编号:1005-0523(2017)06-0066-06

基于粗糙集的道路交通事故客观因素显著性分析

徐 鹏,蒋 凯,王泽华,朱 正

(河海大学土木与交通学院,江苏 南京 210009)

摘要:道路交通事故数据庞大,传统方法对交通事故客观因素的分析较为原始和平面,没有充分挖掘数据之间的潜在联系,且运用传统的方法分析大量交通数据较为困难,所以当今通过数据挖掘技术从大量交通事故数据中挖掘出有联系和价值的信息已成为交通安全领域的热点研究问题。针对传统事故数据统计分析方法的不足,以交通事故数据为基础,建立一种基于粗糙集理论的新型数值模型。通过此数值模型从大量交通事故数据中挖掘分析事故客观因素,以客观因素显著性计算结果为数值依据将各客观因素排序,得出各客观因素对交通事故影响作用大小不同的结论,为以后相关部门交通安全政策的制定提供依据。

关键词:交通安全;数据挖掘;客观因素分析;显著性;粗糙集

中图分类号:U419.3

文献标志码:A

DOI:10.16749/j.cnki.jecjtu.2017.06.008

机动车保有量持续增长且有关驾驶员不遵守交通规则,从而引发的交通事故越发频繁。目前,公安机关已建立庞大的交通事故信息库,而在这些数据中存在着大量的且尚未被发现、存在潜在联系的有价值数据。通过数据挖掘技术对大量交通事故数据进行挖掘分析,已经成为国内外交通安全研究领域的热点问题。然而我国相关部门目前对交通数据库的分析还停留在较为简单、平面和原始的状态,仅仅对事故总起数,人员伤亡数和直接经济财产损失等做宏观统计。本文充分肯定了在交通事故中主观因素的重要性,但交通事故整体特征以及与交通事故发生相关联的客观因素更加值得关注,通过建立数值模型从大量交通事故库中挖掘分析道路交通事故客观因素的显著性,为相关交通部门下一步决策提供依据,从而减少交通事故的发生。

1 国内外研究现状

1.1 国外情况

数据挖掘最早于1989年ACM大会上提出的,是指从大量的数据中提取出隐藏在其中的、人们事先不知道的、潜在的有价值的信息和知识的过程。欧洲环境委员会(EEA)对洲内各个国家的交通的数据情况搜集、分析和挖掘,使得交通管理部门对事故数据的处理变得更加系统^[1]。美国运输研究管理中心通过交通事故数据库,并且与州际高速公路运输协会和公共运输标准委员会等机构进行合作,对国内的汽车安全状况做了翔实的统计分析,并运用数据挖掘的技术对国内的整体汽车安全水平进行了评价^[2]。肯塔基大学运用数据挖掘技术开发出了事故自动的采集系统,总体上达到了事故的数据采集和事故数据处理相对集成化^[3]。日本的工程研究协会运用事故数据和地图建立了与GIS相结合的事故分析系统,对相关数据进行了数据挖掘分析研究,并对北海道的交通事故状况进行了分析。综上所述,欧洲、美国、日本等国家运用先进的数据挖掘技术对交通事故数据进行挖掘分析,除统计基本的参数外,更加注重这些数据之间的联系以及交通事故的原因和影响因素,为相关部门进一步管理和决策提供依据。

收稿日期:2017-07-25

基金项目:江苏省自然科学基金面上项目(BK20151497)

作者简介:徐鹏(1968—),男,副教授,硕士研究生导师,主要研究方向为智能交通与交通环境。

通讯作者:蒋凯(1994—),男,在读硕士,研究方向为智能交通与大数据。

1.2 国内情况

我国与国外在交通事故数据的处理和方案的制定方面还存在很大的差距,我国对事故数据仍然局限于平面分析,仅对事故总起数、伤亡人数、直接经济损失等相关指标做宏观统计,并未运用数据挖掘技术对交通事故数据之间潜在联系进行挖掘研究分析,从而建立起交通事故影响因素之间的联系。从WHO发布的交通事故报告可知,我国以交通事故27.3%的致死率,位居世界之首,可见我国由于缺乏对交通事故的进一步具体分析,挖掘具体的原因,分析隐藏在数据背后的联系,未能及时制定对应的决策方案,成为我国事故死亡率居高不下原因之一。

2 数据挖掘粗糙集方法

数据挖掘是从大量的数据中通过统计、在线分析处理、机器学习等许多方式,根据不完整信息挖掘出与之相关联的信息,揭示隐藏在数据背后的规律,并可以对未来的趋势进行预测^[4]。粗糙集理论已经广泛的应用于其他领域,如机器学习,决策分析等^[5]。

1) 知识表达系统 知识表达系统 S 可表示为 $S=\{U, B, V, F\}$, U 是论域, $B=C \cup D$ 是属性集合, 条件属性 C 与结果属性 D 是子集^[6]。知识表达系统在下面客观因素挖掘分析案例中指的是事故总起数、客观因素和事故类别构成的整体。

2) 上近似和下近似 设 $X \subseteq U$ 是任意一个子集, R 是 U 上的等价关系, 则上近似和下近似集可表示为: $R(X)=\{X \in U, [x]_R \subseteq X\}$, $\bar{R}(X)=\{X \in U, [x]_R \cap X \neq \emptyset\}$ 。上近似和下近似在论文中用于计算属性子集的重要度。

3) 决策表属性依赖度 设 C 和 D 是属性集合, C 对于 D 的依赖度定义为: $\gamma_C = \frac{\text{card}(\text{POSc}(D))}{\text{card}(U)}$, 其中 $\text{POSc}(D)$ 为 D 的 C 正域, $\text{POSc}(D)=\bigcup_{x \in U} CX$, $\text{card}(U)$ 为集合 U 的基数。在实际案例分析研究中指的是事故客观因素集对事故类别属性集的分类相似度。

4) 决策表属性重要度 设 $b \in C$, C 是条件属性集, D 是决策属性集, 则 b 的属性重要度是 $\gamma_C(D) - \gamma_{C-b}(D)$ 。在实际分析案例中指的是某一个客观因素相对于事故类别属性集的显著性。

3 基于粗糙集的交通客观因素挖掘分析

本研究考虑到道路交通事故受许多客观因素的影响,例如道路环境,事故发生时间段,车辆的颜色、型号,车辆品牌,驾驶员的年龄,天气情况,风向和风力,气温等,不同的客观因素对交通事故的影响程度不同,本研究把这种影响程度的差异定义为事故客观因素的显著性。

交通事故客观因素分析有明显的对象性特征,即事故是以对象为基础的,对于不同的对象,事故客观因素显著性也不相同。事故的类别也对道路交通事故客观因素分析产生影响,对不同研究目标,客观分析也可能具有不同显著性^[7]。

道路交通事故客观因素分析的确是基于具体客观因素而言,客观因素显著性也具有相对性,客观因素显著性还需要基于具体的客观因素,各个客观因素显著性之间的比较是相对的。实际的客观因素显著性分析中,客观因素还包括道路环境,车辆情况,驾驶员状态,天气条件,温度等。

基于上述客观因素的特点分析,从而建立起客观因素显著性分析模型

$$\sigma_{C_i}(D)=f(C_i, U, D, C)$$

式中: $\sigma_{C_i}(D)$ 为客观因素 C_i 显著性; C_i 为事故客观因素; U 为事故分析的对象; D 为事故类别属性集合; C 为事故客观因素集合。

在现实的事客观因素显著性分析中,在已知的客观因素分析对象,事故类别属性和客观因素集合前提下,可基于原客观因素显著性分析模型,建立相应的数学模型,用于对客观因素显著性进行数值分析。

粗糙集理论认为不同属性对系统重要性有差异,各属性重要性用重要度来表现。可根据去除该属性,根据系统分类的差别大小来判断该属性的重要性。

假定 $\{U, C \cup D, V, F\}$ 是决策表, 其中 C 和 D 为条件属性集和决策属性集, 属性子集 $C' \subset C$, D 的重要度为 $\sigma_{CD}(C')$, 见下式(1)^[8]:

$$\sigma_{CD}(C') = \frac{|U - BN_R(X)|}{|U|} = \gamma_c(D) - \gamma_{c-c'}(D) \quad (1)$$

式中: $|U|$ 为集合中元素总数; $BN_R(X)$ 为 R 的边界, $BN_R(X) = \bar{R}(X) - \underline{R}(X)$; $\gamma_c(D)$ 为条件属性集 C 对于 D 属性依赖度; $\gamma_{c-c'}(D)$ 为条件属性子集 $C - C'$ 对 D 的属性依赖度。

通过计算去除该条件属性前后属性度依赖度变化的幅度大小来判断该属性的重要程度。属性重要度分析是基于两基本条件的① 明确的决策属性集合② 分析是基于多属性集合而言的。运用粗糙集来分析事故数据也满足这两条基本条件, 所以通过粗糙集理论为基础建立的数值模型来分析交通事故客观因素的显著性具有可行性。

运用原理分析客观因素的显著性, 从而建立起交通事故决策表 $\{U, C \cup D, V, F\}$, 对客观因素 $C_i \subset C$, 由式(2)计算出客观因素 C_i 关于 D 的显著性, 见下式(2)^[9]:

$$\sigma_{Gi}(D) = \gamma_c(D) - \gamma_{c-ci}(D) \quad (2)$$

式中: $\gamma_c(D)$ 为 C 对于 D 分类相似度; $\gamma_{c-ci}(D)$ 为子集 $C - C_i$ 对于 D 分类相似度。

$$\text{其中的 } \gamma_c(D) = \frac{\text{card}(\text{POSc}(D))}{\text{card}(U)}, \gamma_{c-ci}(D) = \frac{\text{card}(\text{POSc}-c_i(D))}{\text{card}(U)}。$$

式中: $\text{card}(\text{POSc}(D))$ 为 C 对 D 正域的基数; $\text{card}(U)$ 为交通事故决策表集 U 的基数; $\text{card}(\text{POSc}-c_i(D))$ 为子集 $C - C_i$ 对于 D 正域的基数。

以上述方法为根据, 可建立起交通事故因素显著性分析的数值模型:

$$\sigma_{Gi}(C') = \frac{\text{card}(\text{POSc}(D))}{\text{card}(U)} - \frac{\text{card}(\text{POSc}-c_i(D))}{\text{card}(U)}$$

通过计算得出交通事故客观因素显著性数值, 以各客观因素数值大小为依托, 对各客观因素进行对比分析。

为了更加精炼的表达出基于粗糙集理论的事故客观因素分析数值模型和方法, 详细的流程图总结如图1所示。

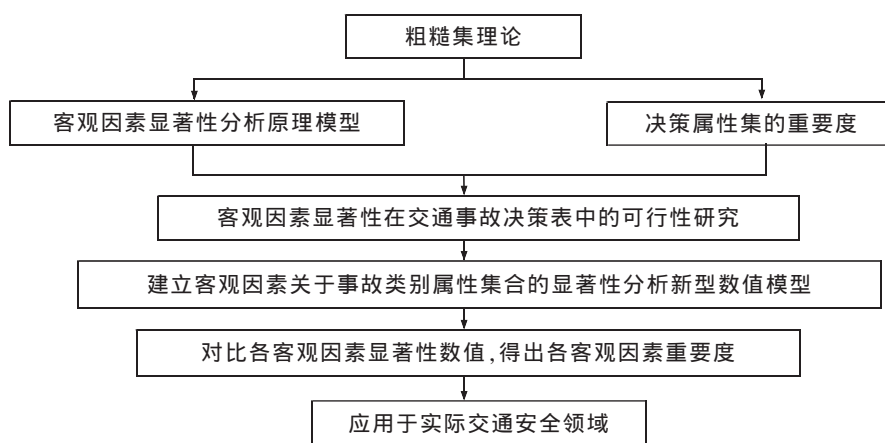


图1 基于粗糙集理论的事故客观因素分析流程图

Fig.1 Analysis flowchart of objective factors based on rough set theory

4 道路交通事故客观因素挖掘分析案例

论文收集了贵阳市 2015 年 1 月 1 日——5 月 31 日的道路交通事故数据, 交通事故数共计 18 041 条。以收集到的数据作为样本, 通过以数值模型来挖掘分析出事故的客观因素显著性。

表1 贵阳市2015年1月至5月交通事故组成
Tab.1 Traffic accident composition in Guiyang City (from January to May, 2015)

事故	追尾	逆行	倒车	未驻车制动	开关车门	违反信号	未按规定让行	其他
数量/起	5 261	69	1 650	176	120	45	7 292	3 428
比例/%	29.16	0.38	9.15	0.98	0.67	0.25	40.41	19.00

选取了事故数据中的7个事故属性作为知识表达系统条件属性 $C, C=\{C_1, C_2, C_3, C_4, C_5, C_6, C_7\}$, 其中 C_1 为发生事故的时间段, C_2 为车辆的颜色, C_3 为车辆型号, C_4 为驾驶员年龄, C_5 为天气条件, C_6 为气温条件, C_7 为风力, 共同组成交通事故客观因素集合。 D 是事故形态决策属性, 它构成了事故类别的属性集合。

C_1 : 1 上午, 2 中午, 3 下午, 4 晚上;

C_2 : 01 白色, 02 银色, 03 黑色, 04 红色, 05 绿色, 06 黄色, 07 灰色, 08 蓝色, 09 其他颜色;

C_3 : 1、别克, 2、奥迪, 3、宝来, 4、宝马, 5、北京现代, 6、比亚迪, 7、宾利, 8、东风标致, 9、东风日产, 10、丰田, 11、福特, 12、大众, 13、其他型号;

C_4 : 1、青年(18-30岁), 2、壮年(30-40), 3、中年(40-50), 4、中老年(50-60), 5、其他;

C_5 : 1、雨, 2、雪, 3、晴, 4、多云, 5、阴天, 6、其他。

C_6 : 1、-1-5度, 2、5-10度, 3、10-15度, 4、15-20度, 5、20-25度, 6、25-30度;

C_7 : 1、风力 ≤ 3 级, 2、其他。

在已经确定了客观因素和类别属性集合后, 然后建立起贵阳市的事故形态知识表达系统, 由于篇幅有限, 仅仅摘录5行数据, 见表2。

表2 交通事故形态知识表达系统
Tab.2 Knowledge system of traffic accident morphology

序号	条件属性 C							决策属性 D
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
1	1	02	1	2	3	1	1	22
2	1	03	12	3	1	2	1	23
3	3	01	13	1	3	3	1	12
4	4	08	5	3	4	4	1	11
5	2	04	9	3	3	3	1	11

在上述的贵阳市道路交通事故形态知识表达系统中, C 是条件属性, $C=\{C_1, C_2, C_3, C_4, C_5, C_6, C_7\}$, D 为决策属性, $D=\{11, 12, 13, 21, 22, 23, 28, 99\}$, 该系统在此挖掘案例分析中表达的是客观因素集决定的事故类别属性集。

U 为论域, $\text{card}(U)=18\ 041$; $\text{POSc}(D)$ 为正域, $\text{card}(\text{POSc}(D))=7\ 412$ 。则代入上述数值可得出客观因素集 C 对类别属性集 D 的分类相似度 $\gamma_c(D)=0.410\ 842$; 为了获取条件属性 C 对决策属性 D 的显著性, 可以假设忽略条件属性 $C_i(i=1, 2, 3, 4, 5, 6, 7)$, 则 $C-C_i$ 对 D 的分类相似度为 $\gamma_{C-C_i}(D)$ 。

计算知识表达系统中条件属性 $\text{card}(\text{POSc}-c_i(D))(i=1, 2, 3, 4, 5, 6, 7)$ 的值, 得 $\text{card}(\text{POSc}-c_1(D))=3\ 756$, $\text{card}(\text{POSc}-c_2(D))=4\ 638$, $\text{card}(\text{POSc}-c_3(D))=4\ 992$, $\text{card}(\text{POSc}-c_4(D))=4\ 168$, $\text{card}(\text{POSc}-c_5(D))=5\ 341$, $\text{card}(\text{POSc}-c_6(D))=5\ 487$, $\text{card}(\text{POSc}-c_7(D))=6\ 459$ 。将属性正域的结果代入公式即可得 $\gamma_{C-C_1}(D)=0.208\ 192$, $\gamma_{C-C_2}(D)=0.257\ 081$, $\gamma_{C-C_3}(D)=0.276\ 703$, $\gamma_{C-C_4}(D)=0.231\ 029$, $\gamma_{C-C_5}(D)=0.296\ 048$, $\gamma_{C-C_6}(D)=0.304\ 141$, $\gamma_{C-C_7}(D)=0.358\ 018$ 。

由公式(2)计算出 $\sigma_{c1}(D)$, $\sigma_{c1}(D)=0.410\ 842-0.208\ 192=0.202\ 65$, $\sigma_{c2}(D)=0.410\ 842-0.257\ 081=0.153\ 761$, $\sigma_{c3}(D)=0.410\ 842-0.276\ 703=0.134\ 139$, $\sigma_{c4}(D)=0.410\ 842-0.231\ 029=0.179\ 813$, $\sigma_{c5}(D)=0.410\ 842-0.296\ 048=0.114\ 794$, $\sigma_{c6}(D)=0.410\ 842-0.304\ 141=0.106\ 701$, $\sigma_{c7}(D)=0.410\ 842-0.358\ 018=0.052\ 824$ 。按道路交通客观因素显著性的大小顺序对上面的 7 个条件属性依次排序: $\sigma_{c1}(D)>\sigma_{c4}(D)>\sigma_{c2}(D)>\sigma_{c3}(D)>\sigma_{c5}(D)>\sigma_{c6}(D)>\sigma_{c7}(D)$ 。

通过上面事故客观因素显著性分析并结合显著性含义,可得到以下结论:

1) 各客观因素对事故影响作用大小有明显的差异,即事故客观因素有主次之分。根据计算所得出的结果,按影响因素从大到小的顺序,将上面 7 个事故客观因素排序如下:发生事故的时间段、驾驶员年龄、车辆的颜色、车辆型号、天气条件、气温条件、风力,如图 2 所示。

2) 各客观因素对交通事故的影响大小与客观因素的取值宽度无明显相关关系,因为客观因素显著性是通过建立显著性分析数值模型,以实际的交通事故数据为基础进行分析的,客观因素显著性与客观因素取值宽度之间不相互影响。各客观因素对交通事故的影响大小依次排序为:发生事故的时间段、驾驶员年龄、车辆的颜色、车辆型号、天气条件、气温条件、风力大小,而与其对应的取值宽度为:7,10,5,4,7,3,7,如下图 3 所示。

3) 通过本道路交通事故客观因素挖掘分析的案例,研究发现,发生事故的时间段和驾驶员的年龄这两种客观因素在交通事故中的影响作用最甚。交通管理部门可以采用限定道路的允许通行时间段的措施,尽可能避开多发事故时间段,将交通事故的发生起数降到最低。考虑到驾驶员年龄对交通事故的影响,必要时相关部门可以重新限定驾驶员的年龄上限或者延长驾照培训考试的学时,这样也有助于减少交通事故的发生。本研究首先充分肯定了在交通事故中人的主观因素是最根本的原因,但客观因素同样不可忽视。因此本研究通过对交通事故大数据进行挖掘分析,为相关部门制定决策提供数据支撑,有一定的应用型和目标性。

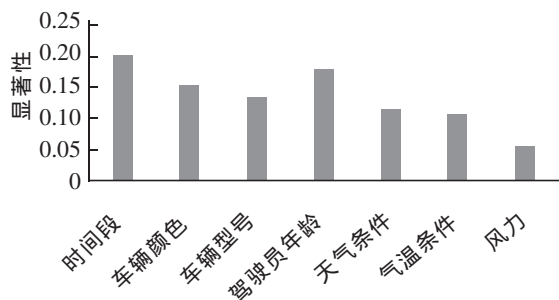


图 2 客观因素对事故影响作用
Fig.2 Effect of objective factors width

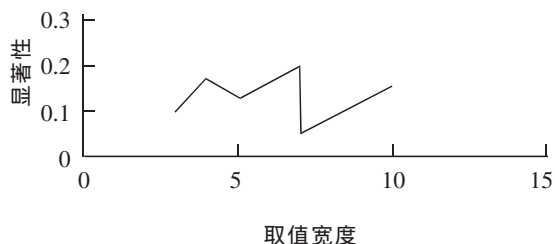


图3 客观因素影响因素与取值宽度
Fig. 3 Influencing factors and value on the incident

5 结论

本研究基于贵阳市的大量交通事故数据,以数据挖掘里面的粗糙集为理论基础,建立一种新型事故客观因素显著性分析实用数值模型,挖掘分析了道路交通事故客观因素的显著性,得出不同客观因素对交通事故影响作用不同的结论,解决了实际交通事故影响因素中具体客观因素的重要度问题。相对于传统的对交通事故数据库的宏观、较为平面的数据分析且将事故的发生简单归结为人的因素,本研究提出的交通事故客观因素显著性分析数值模型,在没有先验知识和事先主观评价的条件下,利用实际的大量交通事故数据,比较客观的从数据中挖掘出其中隐含的规则,从众多客观因素中找出主要的交通事故影响因素,为相关部门作出及时的交通决策方案提供理论事实依据。

考虑到交通事故数据的不公开性,本研究只对简单的算例进行试验,今后还可以选取更多的客观因素,例如道路的线形、照明设施条件、路面的类型等,因此上述模型和算法在交通事故客观因素分析领域的实用性和合理性有待于更全面的交通事故数据的检验。

参考文献:

- [1] DER HORNG LEE, SHIN TING JENG, P CHANDRASEKAR. Applying data mining techniques for traffic incident analysis[J]. Journal of The Institution of Engineers, 2004, 44(2): 90-92.
- [2] C FOREMAN, JR REY, C DEANNUNTIS. National transit bus accident data collection and analysis[J]. Bus Transit Operations, 2002, 81(2): 01-04.
- [3] YANG XUE BING. A aigh effcent multi-dimensional association rules mining algorithm[J]. Computer Development, 2002, 12(6): 52-54.
- [4] 李明祥. 基于粗糙及理论的数据挖掘方法的研究[D]. 济南: 山东科技大学, 2003: 1.
- [5] 任重, 邵军力. 粗糙集理论在通侦信息融合中的应用[J]. 解放军理工大学学报: 自然科学版, 2002, 3(6): 96-99.
- [6] 袁捷. 基于粗糙集的传力杆接缝水泥道面力学响应影响因素分析[J]. 华东交通大学学报, 2017, 34(2): 1-9.
- [7] 程坦. 道路交通事故数据挖掘及应用研究[D]. 哈尔滨: 哈尔滨工业大学, 2009: 33.
- [8] 曾黄麟. 智能计算[M]. 重庆: 重庆大学出版社, 2004: 14-28.
- [9] 陈强, 王双维, 郝乃嫒. 基于粗集理论的交通事故死亡人数时间分布分析[J]. 计算机工程与应用, 2007, 43(8): 217-218.

Analysis of Objective Factors of Road Traffic Accidents Based on Rough Set Theory

Xu Peng, Jiang Kai, Wang Zehua, Zhuzheng

(College of Civil and Transportation Engineering, Hohai University, Nanjing 210009, China)

Abstract: Due to the huge road traffic accident data, it is primitive and flat to use the traditional method of analyzing the objective factors for traffic accidents, which fails to fully exploit the potential link among data. Digging out valuable information from a large number of traffic accident data through the data mining technology has become a hot topic in the field of traffic safety research at present. Aiming at the shortcomings of traditional statistical data analysis methods, this paper, through traffic accident data, established a new numerical model based on the rough set theory. Through this numerical model, the objective factors of the accidents were extracted from the traffic accident data. The objective factors are ranked by the calculation of objective factor significance. The research results show that the effects of objective factors on the traffic accident are different, which may provide some reference for the formulation of the traffic safety policy in the future.

Key words: traffic safety; data mining; objective factor analysis; significance; rough set

(责任编辑 王建华)