

道路交通事故数据深度挖掘技术与应用 ——以深圳市为例

支 野,王大珊,丛浩哲,饶众博
(公安部道路交通安全研究中心,北京 100062)

摘要: 现有道路交通事故统计分析技术存在数据项缺失、方法单一、实战应用性弱等问题,很难为公安交管部门提供针对性的辅助指导。基于深圳市2014—2016年交通事故数据,采用Apriori关联分析算法、贝叶斯理论以及模糊聚类等大数据挖掘方法,探索性地提出道路交通数据缺失数据项填补、事故伤亡特征因子甄别以及事故危险性分类评价方法。结果表明,该方法可有效提高道路交通事故数据完整性和事故伤亡特征因子甄别准确性,以及量化交通事故危险度评价。研究方法和结果可辅助公安交管部门开展道路交通事故预防和交通安全管理工作。

关键词: 道路交通安全; 事故统计; Apriori关联分析; 贝叶斯; K-means聚类

Road Traffic Accident Data Analyzing and Its Application: Example of Shenzhen

Zhi Ye, Wang Dashan, Cong Haozhe, Rao Zhongbo

(Road Traffic Safety Research Center of the Ministry of Public Security, Beijing 100062, China)

Abstract: The existing statistical analysis methods for road traffic accidents is problematic because of missing data items, over simplistic, and weak in applications, which make it hard to be useful for traffic management departments. With the Shenzhen accident data from 2014 to 2016, this paper explosively proposes the methods of Apriori algorithm, Bayesian theory and fuzzy clustering big data mining techniques for solving missing accident attribute data problems, identifying accident severity and classifying accident risk. The results show that these methods can effectively improve the accident data integrity, accuracy of characteristic factor selecting for accident casualties, and the assessment of traffic accident risk quantization. The study methods and results can assist traffic management departments in road traffic accident prevention and traffic safety management.

Keywords: road traffic safety; accident statistics; Apriori correlation analysis; Bayesian theory; K-means clustering

收稿日期: 2018-03-14

基金项目: 国家自然科学基金青年基金项目“多源轨迹数据驱动的城市居民交通出行活动链特征模式研究”(41601434)

作者简介: 支野(1988—),男,陕西宝鸡人,博士,助理研究员,主要研究方向:空间大数据挖掘、智能交通、地理信息系统。E-mail: zhiye007@qq.com

通信作者: 丛浩哲(1982—),男,山东威海人,博士,副研究员,主要研究方向:交通安全工程、交通数据统计分析、交通安全宣传教育。E-mail: conghaozhe@126.com

0 引言

中国正处于机动车、驾驶人及道路里程高速增长期,道路交通事故甚至重特大交通事故时有发生。以2015年道路交通事故统计数据为例,中国道路交通事故数约18.7万起,死亡人数约5.8万人,万车死亡率为2.08,远高于美国(1.25)、日本(0.53)、德国(0.62)等发达国家,中国道路交通安全形势十分严峻^[1]。开展道路交通事故统计分析工

作是预防道路交通事故的有效手段之一。中国每年由公安部交通管理局组织开展道路交通事故统计分析工作,主要是针对交通事故的宏观形势进行统计分析^[2]。

伴随着信息技术步入大数据时代,道路交通事故数据也逐渐呈现出数据量庞大、更新速度快、价值丰富等大数据特征^[3],传统以四项指数(事故起数、死亡人数、受伤人数、直接财产损失)为统计内容、以描述性统计为主要方法的分析手段已难以满足大数

据时代道路交通事故统计分析需要。具体表现在：1)数据项缺失。在实际获取道路交通事故数据过程中，由于多种原因导致数据采集项信息不完整、不准确，致使后续统计分析结果可信度降低。2)方法单一。事故统计方法以描述性统计为主，难以挖掘事故背后深层次的原因和机理，不能量化地甄别事故特征因子及建立事故风险预警评价机制。3)结果不实用。事故数据的挖掘分析多注重方法理论研究，但受数据采集限制，无法获取全项事故数据，因此研究结果一般实用性较单一，没有针对性，也不利于指导基层开展道路交通事故预防工作。

美国、日本、德国等发达国家普遍重视交通事故调查与统计分析工作，设有专门的交通安全研究机构，例如美国国家公路交通安全管理局(National Highway Traffic Safety Administration, NHTSA)、日本交通事故研究分析中心(Traffic Accident Research and Analysis Center, TARAC)、德国联邦公路研究院(Federal Highway Research Institute, BAST)等。这些研究机构普遍建立了道路交通事故分析、预防和评估预警技术体系，统计数据采集全面翔实，较早地将大数据挖掘技术引入事故数据统计分析中，除四项指数等基本统计项外，更加注重挖掘道路交通事故背后潜在的规律和诱因特征^[4-5]。

如何将大数据挖掘技术应用于中国道路交通事故数据分析研判中，为预防事故和降低事故伤亡率提供科学指导，成为新时期公安交管部门面临的巨大挑战。本文以深圳市2014—2016年交通事故数据为研究对象，采用Apriori关联分析算法、贝叶斯理论及模糊聚类等大数据挖掘方法^[6]，探索性地提出道路交通数据缺失数据项填补、事故伤亡特征因子甄别以及事故危险性评价方法。

1 实验数据及预处理

本文获取深圳市2014—2016年交通事故相关数据包括：1)事故描述数据(含事故发生地点信息，见图1)；2)事故涉事人员信息数据；3)路网地图数据；4)以日为单位的天气数据。

经统计，深圳市交通事故信息数据表共有属性项68项，人员信息数据表共有属性项88项。为了更好地开展事故数据深度挖掘，主要进行以下预处理工作：1)数据融合。以事故编码和时空信息为主键，将事故

记录信息、涉事人员信息、路网shp数据以及天气数据进行关联融合，建立事故信息多维度矩阵。2)数据清洗。删除缺失率为100%的属性项，如运载危险品事故后果、是否逃逸等；同时删除对事故诱因深度挖掘关联不大的字段，如调解人、文书状态等字段，最终选取30个字段(见表1)。若发现字段中的属性值不属于《道路交通管理信息代码》(GA/T 16.1—2012)(以下简称《代码》)中所规定的范围，则置为空。3)数据编码。对属性项、分类型属性值进行编码化操作，参照规范建立数据字典表。

2 基于Apriori关联分析的数据填补方法

在实际道路交通事故数据中，由于各种原因导致出现不同程度的数据缺失，其中非伤亡事故比伤亡事故数据缺失率高，直接降低了数据的可用性，不利于数据深度挖掘分析。本文将缺失的事故数据按照是否具有关联性分为两类：一类为随机型缺失数据，表征事故数据中的独立信息，与其他属性项不存在明显的相关性，例如姓名、民族、车牌号码等。该类缺失数据理论上无法通过后期分析进行弥补。另一类为关联型缺失数据，与其他属性值之间存在潜在的关联关系，例如事故形态、交叉口及路段类型、照明条件等，该类缺失数据具有取值范围固定且取值空间相对狭小等特征，可通过大数据挖掘方法进行一定程度的填补，进而提高道路交通事故数据的完整性。

常见的数据填补方法包括：均值填充、热卡填充、聚类填充、多重插补等^[7]。考虑到事故关联性缺失数据以分类数据为主，本文采用关联规则挖掘进行数据填补。关联规则挖掘是数据挖掘的一个重要研究问题，反映一个事物与其他事物之间的相互依赖性 or 相互关联性。关联规则挖掘大量数据项集之间有趣的关联或相关联系，侧重于确定数据中不同领域之间的联系，找出满足给定支持度和置信度阈值的多个域之间的依赖关系。挖掘关联规则是指在数据库中挖掘具有特定形式的规则：由于某些事件(要素)的发生而引起另外一些事件(结果)的发生。关联规则在决策支持系统、专家系统和智能信息系统等方面有着重要的应用价值。

关联规则挖掘过程主要包括三个阶段：第一阶段从数据集中找出所有的频繁项集，均满足支持度 $s \geq \min_sup$ ， $\min_sup$ 为支持

度阈值；第二阶段由这些频繁项集产生关联规则，计算这些关联规则的置信度 c ，然后保留那些满足 $c \geq \text{min_conf}$ 的关联规则， min_conf 为置信度阈值；第三阶段计算关联规则的提升度 l ，并选取提升度 $l > 1$ 的规则。考虑到计算效率，本文选用 Apriori 关联方法进行事故数据潜在关联规则挖掘。具体过程为：1) 设定支持度 $> 3\%$ ，通过 Apriori 算法从事故数据挖掘出频繁项集；2) 设定置信度 $> 60\%$ 、提升度 > 1 为筛选依据，从频繁项集中

形成事故数据关联规则集；3) 按照最大事故信息量原则与最小数据缺失项原则，将具体的事故信息与关联规则集进行类比，获得各项缺失数据的最可能值，对缺失的事故数据进行弥补，进而提高道路交通事故数据的完整性。

以填补伤亡事故中的缺失字段属性值为例，经过 Apriori 关联分析得出强关联且有意义的规则(见表 2)。例如，对于某起伤亡事故，若路表情况字段空缺时，若天气字段为雨天，则可以用规则 1 进行数据填补；若驾龄字段为“小于 5 年”，则可以用规则 3 对数据进行填补，以此类推。经过这种方法处理后，可使事故数据库中的数据进一步丰富，提高数据完整性。

3 基于朴素贝叶斯的道路交通事故伤亡特征因子甄别方法

降低伤亡事故发生频率是公安交管工作的重中之重，目前大多使用数理统计方法对道路交通事故伤亡特征因子进行分析，统计伤亡事故发生概率、分析伤亡事故发生原

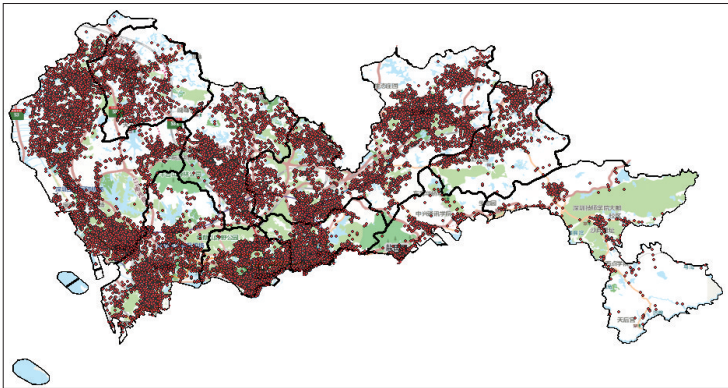


图1 深圳市2014—2016年事故点分布
Fig.1 Distribution of accident locations in Shenzhen from 2014 to 2016

表1 事故数据属性项字典表
Tab.1 Dictionary of accident attribute data

编号	属性项	编码	描述	编号	属性项	编码	描述
1	行政区划	A	1~10分别为深圳市10个区	16	路侧防护设施类型	P	同《代码》第51部分
2	日期	B	1为工作日，2为节假日	17	道路物理隔离	Q	同《代码》第84部分
3	时段	C	1为早晚高峰，2为其他时段	18	路面状况	R	同《代码》第77部分
4	是否有死亡	D	1为有死亡，0为无死亡	19	路表情况	S	同《代码》第78部分
5	是否有受伤	E	1为有受伤，0为无受伤	20	交叉口及路段类型	T	同《代码》第83部分
6	事故认定原因	F	同《代码》第31部分	21	道路线型	U	同《代码》第81部分
7	天气	G	1为晴，2为阴，3为雨，4为雪，5为雾，6为大风，7为沙尘，8为冰雹，9为其他	22	道路类型	V	同《代码》第79部分
8	能见度/m	H	1为<50，2为50~<100，3为100~<200，4为≥200	23	照明条件	W	同《代码》第76部分
9	事故形态	I	同《代码》第43部分	24	公路行政等级	X	同《代码》第80部分
10	地形	J	1为平原，2为山区，3为丘陵，4为其他	25	是否超载	Y	1为无超载，2为超载
11	中央隔离设施	K	同《代码》第74部分	26	车辆类型	Z	同《代码》第4部分
12	性别	L	1为男，2为女	27	驾驶证种类	AA	同《代码》第23部分
13	年龄/岁	M	1为<20，2为20~<40，3为40~<60，4为≥60	28	号牌种类	AB	同《代码》第7部分
14	驾龄/年	N	1为<5，2为5~<10，3为10~<20，4为≥20	29	文化程度	AC	1为未接受过教育，2为小学或初中，3为高中，4为大学本科，5为研究生及以上
15	交通信号控制方式	O	同《代码》第75部分	30	事故责任	AD	同《代码》第48部分

因，以此为基础建立数学模型，来评估道路和交通管理安全性和有效性^[8]。但是这种造成伤亡交通事故的违法行为可能多数情况下导致非伤亡事故的概率更高。例如对深圳市伤亡事故中违法行为统计发现“驾车时有其他妨碍安全行车行为的”约占60%，由此推断只要有这一违法行为就很可能发生伤亡事故是不合适的。因为在非伤亡事故中，这一违法行为也大量出现，其与伤亡事故的比值约为143:1，即发生144次交通事故违法行为为“驾车时有其他妨碍安全行车行为的”时，可能只有一次是伤亡事故。由此可以得出该违法行为作为道路交通事故伤亡特征因子不合适。本文提出基于朴素贝叶斯的故事伤亡特征因子甄别方法：

$$P(S/F_i)=\frac{P(S)\times P(F_i/S)}{P(S)\times P(F_i/S)+P(Q)\times P(F_i/Q)},$$

式中： $P(S/F_i)$ 为出现特征 F_i 时发生伤亡事故的概率； F_i 为第 i 个特征因子； S 为伤亡事故； Q 为非伤亡事故。其中，死亡人数为原始数据中死亡人数与抢救无效死亡人数之和，伤亡率为事故死伤人数/事故涉及当事人数。

经挖掘分析，得出事故发生后造成伤亡率超过80%的前20种伤亡特征因子(见表3)，主要涉及违法行为(5种)、人的因素(2种)、车的因素(5种)、路的因素(6种)和环境因素(2种)。道路交通事故伤亡特征因子能量化地挖掘事故诱因与事故伤亡情况之间

的关联关系，为基层民警开展事故预防预警提供了抓手。

表3 基于贝叶斯的道路交通事故伤亡特征因子甄别
Tab.3 Bayesian theory-based characteristic factor selecting of road accident casualties

特征因子		伤亡率/%	
事故认定原因(违法行为)	驾驶电动自行车超速行驶	96.30	
	机动车在单位院内、居民居住区内不避让行人	93.94	
	行人违反交通信号通行	88.89	
	机动车穿越人行道不减速	87.76	
	行人进入高速公路	82.76	
人	文化程度	未接受过教育	80.40
	驾驶证种类	无证驾驶	86.30
车	车辆类型	轻便二轮摩托车	89.73
		三轮汽车	86.67
		自行车	80.18
	是否超载	超载	92.50
	号牌种类	无号牌	84.50
路	路表情况	有积水	80.00
	中央隔离设施	混凝土护栏	85.19
		波形护栏	88.03
	交叉口及路段类型	高架路段	94.44
		路段出入口	88.89
	桥梁	87.50	
环境	照明情况	夜间无路灯照明	90.69
	交通信号控制方式	无控制	90.90

表2 伤亡事故数据填补规则(部分)
Tab.2 Casualty data filling rules (partial)

编号	填补规则	解释	支持度/%	置信度/%	提升度
1	G3⇒S2	天气：雨⇒路表情况：潮湿	10.0	84.8	5.7
2	E1⇒I11	是否受伤：受伤⇒事故形态：碰撞运动车辆	14.4	62.8	1.2
3	N1⇒M2	驾龄：<5年⇒年龄：20~<40岁	13.3	84.8	1.6
4	H3⇒W1	能见度：100~<200⇒照明条件：白天	9.7	74.7	1.5
5	N1⇒AB2	驾龄：<5年⇒号牌种类：小型汽车号牌	22.7	68.3	3.0
6	L2⇒AD5	驾驶人性别：女⇒事故责任：无责任	11.8	60.2	1.8
7	O7 ⇒I11	交通信号控制方式：无控制⇒事故形态：碰撞运动车辆	18.4	70.8	1.4
8	AD3 ⇒E0	事故责任：同等⇒是否有受伤：无受伤	21.9	83.0	1.3
9	N2⇒L1	驾龄：5~<10⇒驾驶人性别：男	28.9	97.7	1.2
10	Q4⇒ I11	道路物理隔离：中央隔离加机非隔离⇒事故形态：碰撞运动车辆	33.6	61.4	1.2
.....					

注：本文选取的都是单维关联规则，多维关联未考虑。

4 基于聚类算法的道路交通事故危险性挖掘

道路交通事故起因复杂多样,每起事故都有其独特的特点,但是在多个事故中可能存在相同的特征,本文称之为共性特征。掌握交通事故的共性特征、发现交通事故危险程度的规律性结论,有助于在警力资源有限的情况下,有针对性地采取预防措施,减少交通事故的发生。本文基于K-means聚类算法,以事故认定主要违法行为为对象,以事故数、伤亡率(基于表3中结果)、受伤人数、死亡人数作为危险度分析指标,对事故数据进行聚类。算法过程为:1)建立事故认定主要违法行为的分析矩阵,主要违法行为为行,4个危险度指标为列;2)采用SVD算法对矩阵进行降维分解,选定前 n 个分量作为特征分量(本文选取前2个,见图2a);3)选取 k 个初始聚类中心;4)确定相似性距离度量函数。采用欧式距离作为计算两两违法行为前 n 个特征分量的相似性距离。5)开始迭代更新运算。计算每个样本数据与每个聚类中心的相似性距离,将样本归到距离最短的

类中。6)更新聚类中心位置。利用均值方法,更新每类的中心点位置。重复5)运算,直至每个聚类中心值保持不变。

本文基于深圳市事故数据进行聚类分析后,得到3个具有明显特征的类簇(见图2)。类别1主要是事故频次高、伤亡率较低的事故记录,包括变更车道时影响正常行驶、不按规定倒车、不让右行等发生在城区的交通事故。类别2主要是事故频次较高、伤亡率较高的事故记录,包括不按规定会车、转弯机动车未让直行车辆、行人先行,驾车时有其他妨碍安全行车行为等涉及机动车的事故。类别3主要是事故频次较低、伤亡率高的事故记录,包括未按规定戴安全头盔、驾驶电动自行车超速行驶、肇事逃逸构成犯罪等以摩托车、电动车、非机动车为主的事故。

道路交通事故危险度分析结果具有重要的应用价值:一方面,可以对特定道路、交通、环境条件下各类交通事故的危害性进行量化,便于相互之间比较;另一方面,事故危险性可作为该类事故的权重纳入道路交通安全评价中,提升评价结果的合理性。

5 结语

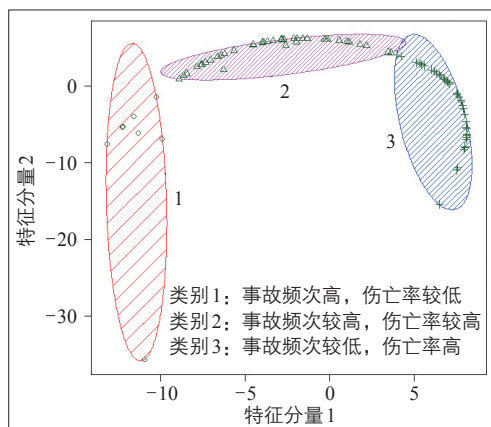
本文针对现有道路交通事故统计工作中存在的数据项缺失、方法单一、实战应用性弱等问题,借助大数据挖掘技术,探索性地提出道路交通数据缺失数据项填补、事故伤亡特征因子甄别,以及事故风险等级分类评价方法,挖掘出的结果可辅助公安交管部门开展道路交通事故预防和交通安全管理工作。本文研究成果仍存在一定的局限性,例如时段仅划分了早晚高峰和平峰,未考虑白天和夜间的区别,天气未考虑雨量和风力。另外,考虑到深圳城市发展与交通管理政策,并未对深圳特区内外分别进行研究。未来,将进一步对道路交通事故数据开展多角度、精细化深入挖掘,同时将研究道路交通事故数据挖掘的自动化和智能化,研发相关决策支持系统。

参考文献:

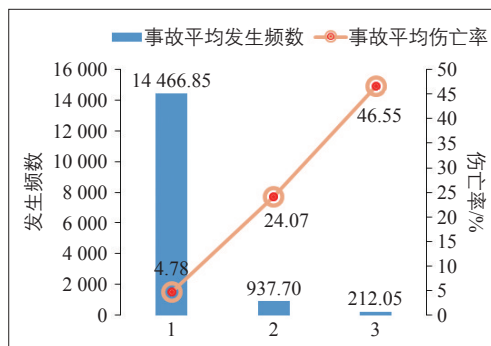
References:

- [1] 公安部交通管理局. 中华人民共和国道路交通事故统计年报(2015年度)[R]. 北京: 公安部交通管理科学研究所, 2016.

(下转第61页)



a 事故危险性特征聚类



b 三种类型事故特征

图2 基于事故危险性特征的聚类结果

Fig.2 Clustering results based on accident risk characteristics