粗集在交通事故黑点成因分析中的应用

张 鹏[1]，张 靖[2]，刘玉增[1]，唐雪飞[1]

(1. 电子科技大学计算机科学与工程学院 成都 610054; 2. 攀枝花学院网络中心 四川 攀枝花 617000)

【摘要】交通事故黑点的形成原因多种多样，每一个事故多发点都有其形成的最大诱因。作为交通安全管理工作的重要任务，交通事故黑点的鉴别与改善是预防交通事故发生、减少交通事故损失的有效手段。该文针对交通事故多发点成因的复杂性和多样性，提出通过粗集来对公路交通中的不利因素进行筛选，找到形成事故多发点的最大诱因，从而有针对性地进行整治，能够有效地节约时间和费用，避免不必要的人力、物力浪费。

关 键 词 决策表; 知识库; 粗集; 交通黑点
中图分类号 TP331 文献标识码 A

Rough Set Application in the Analysis of Formulation Cause of Traffic Black-Spot

ZHANG Peng[1]，Zhang Jing[2]，LIU Yu-zeng[1]，TANG Xue-fei[1]

(1. School of Computer Science and Engineering, Univ. of Electron. Sci. & Tech. of China Chengdu 610054;
2. Campus Network Center, Panzhihua University Panzhihua Sichuan 617000)

Abstract The traffic black-spot cause of formation is very complex, and every black-spot has the most important inducement for itself. The identification and improvement of black-spot, as an important task of traffic safety management, are very effect measures to reduce the traffic accident frequency. For the complexity and diversity of the black-spot cause of formation, we find a method based on the rough set to determine the most important inducement of the black-spot, then we can repair the inducement in effect and save more time and money.

Key words decision-making table; knowledge base; rough set; traffic black-spot

交通事故黑点即交通事故多发点的形成存在着许多复杂的因素，且这些因素对事故结果具有不同程度的影响。如何对这些因素进行分类，区分主次要因素，是决定事故多发点整治方案的首要问题。以往人们往往通过先验经验或专家经验来决定其主要因素，提出解决方案。但这种方法存在很大的随意性和主观性，一旦导致事故的真正原因没有找到，就会造成该事故多发点处的交通问题不能解决，投入整治的资金白白浪费。

粗集理论是一种新的处理模糊和不确定知识的数学工具，其主要思想是在保持信息系统的分类能力不变的前提下，通过知识约简，导出问题的决策或分类规则。粗集理论的优势在于它不需要先验知识，便可完全从数据或经验中获取知识生成决策规则。利用粗集可以对属性的重要性进行度量，这个度量是根据论域中的样例来得到的。根据得出的属性重要性，可以确定各个属性的权重，从而判断出某个事故多发点形成的主要因素和次要因素等，进而有针对性地提出整治方案。

1 粗集的相关知识

设 U 是所论述个体的全域，存在子集 $X \subset U$，定义 R 为一等价关系，当 X 能用 R 属性集确切地描述时，它可用某些 R 基本集合的并来表达，称 X 是 R 可定义的，否则 X 为 R 不可定义的。R 可定义集也称作 R 精确集，而 R 不可定义集则称为 R 非精确集或 R 粗集。对粗集可使用两个精确集，即粗集的上近似集和下近似集来描述，也可对粗集近似地定义。考虑两个子集[1]：

$$R_{-}(X)=\{X \subset U:[X]_R \subset X\} \tag{1}$$

$$R^{-}(X)=\{X \subset U:[X]_R \bigcap X \neq \varnothing\} \tag{2}$$

分别称它们为 X 的下近似集和上近似集。

集合 $\mathrm{POS}_R(X)$ 称为 X 的 R 正域，定义如下：

$$\mathrm{POS}_R(X)=R_{-}(X) \tag{3}$$

收稿日期：2005－03－16
作者简介：张 鹏(1979－)，男，硕士，主要从事软件工程与算法设计等方面的研究.

$R\_(X)$ 是根据知识 R、U 中所有一定能归入 X 的元素的集合。$R^{-}(X)$ 是根据知识 R、U 中一定能和可能归入 X 的元素的集合。可用 $U|R$ 表示根据关系 R、U 中的对象构成的所有等价类族。对于近似分类的不精确性的度量也可以定义为[6]：

$$r_R(F)=\frac{\sum_{i=1}^{n}\mathrm{card}(R\_(X_i))}{\mathrm{card}(U)} \tag{4}$$

式中 card( )表示集合中元素的数目。

2 事故数据预处理

交通事故中收集到的原始数据不一定能直接用于知识获取，通常还需要进行预处理加工。对于原始数据资料中遗漏的信息，需要补充(在基于粗集理论的知识获取中称为决策表补齐)。对于原始资料中值域为实数值的数据，还需要进行离散化，因为粗集理论研究的元素对象只能是离散化对象。

2.1 决策表的补齐

交通事故多发点数据决策表是进行数据处理、得出事故结论的基本数据结构，是一类特殊而重要的知识表达系统(KRS)。知识表达系统的基本成分是研究对象的集合，关于这些对象的知识可通过指定对象的基本特征(属性)和它的特征值(属性值)来描述。一个知识表达系统可表述为$S=(U, C, D, V, f)$，其中 U 是对象的集合；$C\cup D=R$ 是属性的集合；子集 C 和 D 分别称为条件属性集和结果属性集；$V=\bigcup_{r\in R}V_r$ 是属性值的集合，V_r 表示了属性 $r\in R$ 的属性范围；$f:U\times R\to V$ 是一个信息函数，它指定 U 中每一对象 x 的属性值。这样定义的知识表达系统可以方便地用表格表达来实现。

定义交通事故多发点数据决策表为$S=(U, A, V, f)$，其中 $A=\{a_1,a_2,\cdots,a_n\}$。A 中的元素表示造成交通事故的各项可能因素，用 a_i 表示。$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。

定义 $\boldsymbol{M}(i, j)$ 表示经过扩充的可辨识矩阵中第i行第j列的元素，则经过扩充的可辨识矩阵为[3]：

$$\boldsymbol{M}(i,j)=\{a_k\mid a_k\in A\wedge a_k(x_i)\neq a_k(x_j),\wedge a_k(x_i)\neq *,\wedge a_k(x_j)\neq *\} \tag{5}$$

式中 $i, j=1,2,\cdots,n$；“ * ”为空缺属性值(遗失值)。

为了使具有遗失值的对象与信息系统的其他相似对象的属性值尽可能保持一致，即使属性值之间的差异尽可能保持最小。本文采用不完备数据分析算法ROUSTIDA，它是以可辨识矩阵为基础的。算法ROUSTIDA参见文献[3]。

2.2 决策表的离散化

如果把属性值的定性和定量描述都称为连续值，由于粗集不能运用在具有连续变量的数据上，则必须先对决策表进行离散化。一个解决的办法就是把这些连续的变量分割成一系列区间，落在每个区间内的数据都看成是相等的，这种方法叫作离散化。离散化方法应满足[4]：(1) 属性离散化后的空间维数尽量小，也就是每一离散化后的属性值的种类尽量少；(2) 属性值被离散化后的信息丢失尽量少。而寻求最优的离散化结果已经被证明是NP完全问题。本文采用一种全局聚类分析方法[6]来实现属性值的离散化处理。

令论域 U 中研究的对象数目为 m；条件属性值 $C=\{c_1,c_2,\cdots,c_n\}$；令 x_i 是论域 U 中研究的对象，则：

$$x_i=\{c_{i1},c_{i2},\cdots,c_{in}\}\qquad i=1,2,\cdots,m \tag{6}$$

式中 $c_{i1},c_{i2},\cdots,c_{in}$ 是 x_i 的 n 个属性的连续属性值。

为了将这些属性值通过聚类分析来离散化，首先要将相近似的对象找出来，把它们分为一类，然后把这些类的属性值对应分组。对于每一属性，每种分组，其值相近似的又合并为一个区域，每一个区域一个代码，即属性值属于该区域的值时，就认为该属性的离散化值等于该区域的代码。具体步骤如下：

对论域 U 中的对象，从 $i=1\sim m$，计算每两个对象的Euclidean距离：

$$d_{ij}=\sqrt{(a_{i1}-a_{j1})^2+(a_{i2}-a_{j2})^2+\cdots+(a_{in}-a_{jn})^2}$$
$$i\neq j, i=1,2,\cdots,m, j=1,2,\cdots,m \tag{7}$$

(1) 如果某一 d_{ij} 最小，就认为该 x_i 与 x_j 构成了一个新类。由构成的新类再与论域中其他的对象计算距离。如对于聚类 a 和由聚类 b 和 c 构成的一个新的聚类 bc 的距离：

$$d_{a(bc)}=\frac{(d_{ab}+d_{ac}+d_{bc})/2}{2} \tag{8}$$

(2) 重复进行上述过程，每一次聚类产生一种新的划分，直到划分的协调度等于或大于原始数据的协调度。协调度可以按下式计算：

$$L_d=\frac{\sum_{x\in\{d\}}\mathrm{card}(C\_X)}{\mathrm{card}(U)} \tag{9}$$

式中 d 为决策属性的全体等价类集合，新的划分与原始划分 d 显然不同；C_X 为 X 的下近似集合。

(3) 根据最后的聚类，如 k 类，对每一属性 c_j，得到 k 个属性值聚类，即：

$$P_{xcj}=\{c_{xj}, x\in k, j=1,2,\cdots,n\}\qquad x=1,2,\cdots,k \tag{10}$$

(4) 对于每一属性 c_j，每一聚类定义一个属性值区间为：

$$I_j = [c_{xj_{\min}}, c_{xj_{\max}}] \tag{11}$$

(5) 合并 c_j 相似的属性值的区间，并对合并后的区间重新编码。如果连续属性值属于那一个区间的值时，其离散化值就等于该区间的代码。

3　黑点成因分析

利用决策表中的数据，可以计算出各个条件属性的影响强度，影响强度是用属性重要性来进行评判的。

3.1　属性重要性的判断

为了找出某些属性或属性集的重要性，本文的方法是从决策表中去掉一些属性，再来考察没有该属性后决策表中的分类会怎样变化。若去掉该属性会相应地改变分类，则说明该属性的影响强度大，即重要性高；反之说明该属性的影响强度小，即重要性低。对于属性集 D 导出的分类的属性子集 B' 包含于 B 的重要性，用两者依赖程度的差来量度，即：

$$r_B(D) - r_{B-B'}(D) \tag{12}$$

这表示从集合 B 中去掉某些属性子集 B' 后，再对对象进行分类时，结果属性集的分类 U/D 的正域将会受到怎样的影响。

3.2　应用实例

表1　某路段交通事故统计表

	x_1	x_2	x_3	x_4	x_5
a_1	中速	直道	气候多变	死亡0人	警告
a_2	低速	连续弯道	气候平常	死亡0人	危险
a_3	低速	弯道	气候恶劣	死亡2人	危险
a_4	中速	弯道	气候多变	死亡2人	警告
a_5	低速	直道	气候多变	死亡1人	极度危险
a_6	超速	弯道	气候恶劣	死亡2人	危险
a_7	中速	连续弯道	气候恶劣	死亡0人	危险
a_8	超速	连续弯道	气候平常	死亡1人	极度危险

本文用一个如表1所示的某路段交通事故统计表来作为示例。考虑到交通事故的发生和季节、气候的关系十分密切，用集合 $A=\{a_i \mid i=1,2,\cdots\}$ 表示事故的集合，a_i 表示某一个具体事故；用$X=\{x_i \mid i = 1, 2, \cdots\}$表示事故成因的集合，具体设定如下：$x_1$ = 车辆速度(0表示低速，1表示中速，2表示超速)；x_2 = 道路几何线形(0表示直道，1表示弯道，2表示连续弯道)；x_3 = 气候状况(0表示气候平常，1表示气候变化较大，2表示气候恶劣)；x_4 = 伤亡情况(0表示无死亡事故发生，1表示一年内总共死亡人数大于0人小于3人，2表示一年内总共死亡人数大于等于3人)；x_5 = 黑度，黑度是指某一交通事故黑点的危险程度，越危险则黑度越高。本文将导致事故的因素的危险程度分为三级，用{0，1，2}分别表示{警告，危险，极度危险}；$a_i(x_j)$ 的值是第j个因素在第i个具体事故中的值。表2是表1离散化后的结果。

表2　离散化表1

	x_1	x_2	x_3	x_4	x_5
a_1	1	0	1	0	0
a_2	0	2	0	0	1
a_3	0	1	2	1	1
a_4	1	1	1	1	0
a_5	0	0	1	1	2
a_6	2	1	2	1	1
a_7	1	2	2	0	1
a_8	2	2	0	1	2

下面计算表2中属性的重要性。属性集：

$$C=\{x_1,x_2,x_3\}, D=\{x_4,x_5\}$$

则：

$$U/C=\{\{a_1\},\{a_2\},\{a_3\},\{a_4\},\{a_5\},\{a_6\},\{a_7\},\{a_8\}\}$$
$$U/D=\{\{a_1\},\{a_2,a_7\},\{a_3,a_6\},\{a_4\},\{a_5,a_8\}\}$$

得到：

$$\mathrm{POS}_C(D)=\{a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8\}$$

进而得到属性D相对于属性C的依赖性：

$$r_C(D)=8/8=1$$

下面计算属性 x_1,x_2,x_3 相对于属性 x_4,x_5 的重要性。由于：

$$U/\{x_2,x_3\}=\{\{a_1,a_5\},\{a_2,a_8\},\{a_3,a_6\},\{a_4\},\{a_7\}\}$$
$$U/\{x_1,x_3\}=\{\{a_1,a_4\},\{a_2\},\{a_3\},\{a_5\},\{a_6\},\{a_7\},\{a_8\}\}$$
$$U/\{x_1,x_2\}=\{\{a_1\},\{a_2\},\{a_3\},\{a_4\},\{a_5\},\{a_6\},\{a_7\},\{a_8\}\}$$

则：

$$\mathrm{POS}_{C-x_1}(D)=\{a_3,a_4,a_6,a_7\}$$
$$\mathrm{POS}_{C-x_2}(D)=\{a_2,a_3,a_5,a_6,a_7,a_8\}$$
$$\mathrm{POS}_{C-x_3}(D)=\{a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8\}$$

即：

$$r_{C-x_1}(D)=4/8=0.5$$
$$r_{C-x_2}(D)=6/8=0.75$$
$$r_{C-x_3}(D)=8/8=1$$

因此：

$$r_C(D)-r_{C-x_1}(D)=1-0.5=0.5$$
$$r_C(D)-r_{C-x_2}(D)=1-0.75=0.25$$
$$r_C(D)-r_{C-x_3}(D)=1-1=0$$

由此可见，结果属性对条件属性 x_1 即车辆速度的依赖程度最大，即条件属性 x_1 的影响强度最大，

条件属性 x_2 即道路几何线形的影响强度次之。车辆速度是否合理是导致该路段事故发生的最大诱因，而道路的几何线形是其次的诱因。针对这两个条件，交通管理部门就可以制定出相应的整治措施，如在适当的路段对经过车辆的最低速度和最高速度作出限制，在弯道处设置警示牌，安置凸面镜以方便司机观察弯道对面来车情况等。

4 结 束 语

本文通过运用粗集来对形成交通事故黑点的诸多因素进行筛选，能准确合理地分析出产生交通黑点的主要因素和一般因素，从而能针对其主要因素进行整治，节省了治理交通黑点的时间和费用，有效地避免了人力物力的浪费。

参 考 文 献

[1] PAWLAK Z. Rough sets[J]. Communication of ACM, 1995, 38(11): 89-95.
[2] 陈世清, 唐志航, 肖建华. 基于粗糙集联系度的数据挖掘算法及应用研究[J]. 计算机应用, 2004, 24(6): 74-75.
[3] 王国胤. Rough集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.
[4] 曾黄麟. 粗集理论及其应用[M]. 重庆: 重庆大学出版社, 1998.
[5] 王德松, 舒 兰. 粗集决策表与决策表简化的可信度比较[J]. 电子科技大学学报, 2004, 33(5): 611-612.
[6] 曾黄麟 智能计算–关于粗集理论、模糊逻辑、神经网络的理论及其应用[M]. 重庆: 重庆大学出版社, 2004.

编 辑 漆 蓉

(上接第241页)

4 结 论

出于对系统的安全性和用户使用的方便性考虑，需要保证分布式并行安全操作系统的各个节点的用户信息的一致性。为了达到这个目的，本文提出了一种解决方案，并且予以实现。实验证明：该方案能够有效地对分布式并行安全操作系统的各个节点的用户信息进行全局同步，并且能保证在任意节点上执行用户操作后用户信息及在遇到节点故障或网络故障之后恢复各个节点用户信息的一致性。

参 考 文 献

[1] GALLI D L. 分布式操作系统: 原理与实践[M]. 徐良贤, 译. 北京: 机械工业出版社, 2003.
[2] 刘详岐, 卢显良. 分布式集成防御系统中的任务迁移[J]. 实验科学与技术, 2006, 4(3): 28-30.
[3] SANDBU R S, COYNE E J, FEINSTEIN H L, et al. Role-based access control models[J]. IEEE Computer, 1996, 29(2): 38-47.
[4] SMALLEY S, VANCE C, SALAMON W. Implementing SELinux as a Linux security module[EB/OL]. http://www.nsa.gov/selinux/papers/module.pdf, 2004-11-05.
[5] ZHANG C N, YANG Cun-gang. An object-oriented RBAC model for distributed System[C]// Proceedings of the Working IEEE/IFIP Conference on Software Architecture (WICSA'01). Amsterdam Netherlands: IEEE Computer Society Washington, 2001, 08: 24-32.
[6] ZADOK E. Linux网络文件系统管理指南[M]. 邱仲潘, 译. 北京: 电子工业出版社, 2001.
[7] SCHNEIER B. 网络信息安全的真像[M]. 吴世忠, 译. 北京: 机械工业出版社, 2001.

编 辑 漆 蓉