基于 FCM 与模糊粗糙集理论的交通事件检测模型

张慧哲,王 坚,梅宏标
ZHANG Hui-zhe,WANG Jian,MEI Hong-biao

同济大学 CIMS 研究中心,上海 201804
CIMS Research Center,Tongji University,Shanghai 201804,China
E-mail:zhanghuizhe168@163.com

ZHANG Hui-zhe,WANG Jian,MEI Hong-biao.Traffic incident detection model based on FCM and fuzzy rough sets theory.Computer Engineering and Applications,2008,44(23):4-7.

Abstract: In order to detect traffic incident accurately,reduce traffic delays and ensure road safety,a new detection model using FCM and fuzzy rough sets is presented.The model is composed by three parts such as discretization,reasoning rules establishment and fuzzy reasoning.A new method is proposed when attribute discretization.Using custom membership function to fit results of the FCM clustering,the membership function and parameters are obtained to discrete the attribute data without clustering when new data come.Then reasoning rules are formed using the rough set theory so as to improve fuzzy reasoning.Finally Max-Min fuzzy reasoning method is adopted to incident detection.By comparative testing,the better performance is achieved using new model and result valid the validity of new model which provides a new idea for automatic incident detection.

Key words: incident detection;fuzzy C mean clustering;rough sets;fuzzy inference;attribute discretization

摘 要:为准确及时地发现高速公路上的事故隐患,有效地减少交通延误,保障道路安全,提出了一种新的基于模糊 C 均值(FCM)聚类和模糊粗糙集的交通事件自动检测模型。模型分为离散化、推理规则建立和模糊推理三个步骤。在属性离散化时,提出用常用的隶属度函数来拟合 FCM 聚类后的结果,并用此函数和参数来实现属性数据的离散化,避免了每次输入数据都必须通过聚类操作来进行离散化;采用了粗糙集理论建立推理规则,选择和交通事件密切相关属性并进行规则的约简,加速了模糊推理的速度;最后采用 Max-Min 模糊推理方法对交通事件进行检测。通过多种检测方法对比测试,结果表明了此模型在总体性能上优于传统的检测方法,验证了此模型的有效性,为交通事件的检测提供了一种新的思路。

关键词:事件检测;模糊 C 均值聚类;粗糙集;模糊推理;属性离散化
DOI:10.3778/j.issn.1002-8331.2008.23.002 **文章编号**:1002-8331(2008)23-0004-04 **文献标识码**:A **中图分类号**:TP311.13

1 引言

高速公路事件包括交通事故、车辆故障、货物散落、道路维修等破坏正常交通流并导致通行能力下降的事件,是造成交通延误的最主要原因。2008 年席卷中国南部的雪灾与冰冻使人们又一次认识到,如何快速准确地检测高速公路交通事件对减少交通延误、保障道路安全、避免二次事故的发生、减少人身和财产的损失等具有十分重要的意义,建立先进的事件检测系统是高速公路运营管理中的一项首要工作。

目前普遍应用的各种交通事件自动检测算法大致可以归纳为两大类,直接检测算法和间接检测算法。直接检测算法如视频检测法[1],通过摄像机拍摄视频,然后通过图像采集卡从实时视频采集到图像序列,然后对图像进行分析处理,对车辆进行检测、跟踪和速度估算,在统计交通流的同时对高速公路上的事件进行检测。该方法可以很好地从微观角度对事件性质进行检测,其主要缺点是需要密集地安装摄像机,成本较高,且受天气影响较大。间接检测算法是根据事件对交通流的影响来检测事件的存在,大多数间接检测算法是通过对主线上设置的检测器采集到的交通流参数分析判断是否有事件发生,该类检测算法主要有加利福尼亚算法[2]、贝叶斯算法、McMaster 算法[3]、时间序列算法[4]、小波分析算法[5]等等,这些算法具有成本低,简单易操作等特点,但是存在检测率不高,误报率较大等问题,人工神经网络[6]是一种智能的间接检测方法,具有检测率较高,误

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2003AA414120);国家科技支撑计划项目(Project Supported by the National Technology Plan,China No.2006BAF01A46);上海市社会发展重大专项项目(Important Special Project of Social Development,Shanghai No.06DZ12001);上海市基础研究重点项目(Foundation Research Significant Project,Shanghai No.06JC14066);上海市科技发展基金重点项目(Significant Project Supported by Technology Development,Shanghai No.061612058);上海市登山行动计划项目(Project Supported by Climbing-hill Plan,Shanghai No.061111006)。

作者简介:张慧哲(1979-),女,博士研究生,研究方向:数据仓库、数据挖掘及智能交通系统;王坚(1961-),男,研究员,博士生导师,主要研究方向:先进制造技术、企业信息化、数据仓库技术;梅宏标(1976-),男,博士研究生,研究方向:智能交通系统、分布式仿真。

收稿日期:2008-03-31 **修回日期**:2008-05-13

报率较低等特点，但是人工神经网络结构的确定没有一个固定的准则，需要大量的训练样本，且存在着过拟合现象，容易陷入局部最小值等问题，模型泛化能力不强。

本文在已有研究成果的基础上提出了基于 FCM 和模糊粗糙集的检测方法。通常影响交通状况的因素较多，如果直接采用模糊推理会导致模糊推理规则成倍的增加，导致对检测的效率造成一定的影响。本文首先利用聚类算法实现了对输入空间的划分，避免了隶属函数确定的盲目性；通过使用常用的隶属度函数来拟合聚类后得到的结果，取得隶属度函数的参数，从而对属性数据进行离散化，避免了每次输入数据都通过聚类来进行离散化的弊端；并利用粗糙集进行属性约简解决了模糊推理规则“组合爆炸”的问题。

2　模糊粗糙集的交通事件检测模型

本文所提出的交通事件检测模型如图 1 所示。

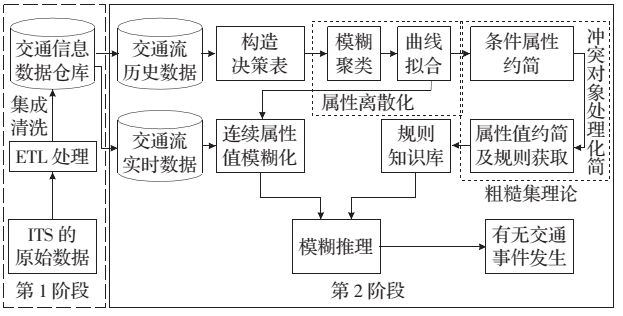

图 1　交通事件的检测模型

交通事件的检测过程分为两个阶段：第 1 阶段为了加速在线检测，构造了交通信息数据仓库，将数据经过 ETL（抽取、转换、加载）及数据融合和清洗后为交通事件的发现提供一个干净、一致、集成、归约（reduction）的数据集；第 2 阶段是建立交通流检测模型，即从交通数据仓库中分别提取出交通流历史数据和实时交通流数据集市，执行不同的操作处理，最终实现交通事件的检测。

第 2 阶段的交通流检测模型处理主要包括以下几个主要步骤：

（1）构造决策表。从数据仓库中取出影响交通状况的相关因素作为条件属性，观测者或管理者对交通状况的评价作为决策属性，形成一张二维决策表格，表中每一行描述一个对象，每一列表征对象的一种属性；

（2）属性离散化。为使量化后的决策表具有最大一致性，笔者应用模糊聚类方法对样本集进行聚类，并对聚类的结果进行曲线的拟合，从而实现了样本的离散化；

（3）粗糙集约简。将决策表离散化后，应用粗糙集对决策表进行简化，主要包括各条件属性和属性值的简化；

（4）规则知识库的建立。模糊理论是处理系统不确定性的重要工具，在日常生活中常存在流量、速度、占有率等的不确定性描述，应用本文后面所描述的模糊离散化方法和各推理规则的约简结果，建立推理规则集；

（5）模糊推理。这里采用应用 Max-Min 模糊推理方法，从而得出交通事件的状态。

3　连续属性离散化及模糊化

3.1　FCM 聚类算法原理

在基于粗糙集理论的交通事件规则提取中，要求决策表中的值用离散数据（如 Integer、String 等）表达，交通流数据有必要对其进行离散化处理。目前，对各个连续数据进行离散化的主要方法有等宽区间法、等频区间法、统计法、滤波法和遗传算法等。连续属性的离散就是将连续属性值域划分为若干区间，每个区间用不同代码表示，这样连续属性值便转换为离散属性值。这里采用了模糊聚类的算法对样本空间进行模糊聚类，由于其适应性强、操作简便，已在图像分割、语音识别等多个领域广泛应用[10]。

模糊 C 均值算法的聚类准则函数为：

$$J_m=\sum_{j=1}^{N}\sum_{i=1}^{c}(\mu_{ij})^m\|x^j-\omega_i\|^2$$

其中：x^j 为样本空间数据 $j=1,2,\cdots,N$；ω_i 为聚类中心，$i=1,2,\cdots,c$；μ_{ij} 为 x^j 对 ω_i 的隶属度，且满足 $\sum_{i=1}^{c}(\mu_{ij})=1$，$\mu_{ij}\in(0,1)$，$m\in(1,\infty)$ 为权重指数，通常取值为 2。

隶属度计算公式为：

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left[\frac{\|x^j-\omega_i\|^2}{\|x^j-\omega_k\|^2}\right]^{1/(m-1)}},i=1,\cdots,c,j=1,\cdots,N$$

聚类中心计算公式为：

$$\omega_i=\frac{\sum_{k=1}^{c}(\mu_{ij})^m x^k}{\sum_{k=1}^{c}(\mu_{ik})^m},i=1,\cdots,c$$

FCM 的过程就是最小化 J_m 的过程，算法步骤如下：

步骤 1　给定类别数 c，参数 m，容许误差 ξ 的值；

步骤 2　随机初始化聚类中心 ω_i，$i=1,2,\cdots,C$，并令循环次数 $k=1$；

步骤 3　按隶属度计算式 $\mu_{ij}(k)$，$i=1,\cdots,c,j=1,\cdots,N$；

步骤 4　按式（3）修正所有的聚类中心 $\omega_i(k+1)$，$i=1,\cdots,c$；

步骤 5　计算误差 $e=\sum_{i=1}^{c}\|\omega_i(k+1)-\omega_i(k)\|^2$，如果 $e<\xi$，算法结束；否则 $k=k+1$，转步骤 3；

步骤 6　样本归类，算法结束后，可按下列方法将所有样本归类：$\|x^j-\omega_i\|^2<\|x^j-\omega_k\|^2$，$k=1,2,\cdots,C,k\neq i$，则将 x^j 归入第 i 类。

在聚类操作中，若对每一个输入数据都通过聚类来进行模糊化处理，一方面由于聚类速度比较慢，会严重影响整个系统的性能；另一方面，由于每次聚类所用的数据集有所不同，聚类后得到的聚类中心也会有所不同，会导致系统结构不稳定。因此，本文提出通过常用的隶属度函数来拟合聚类后得到的结果，取得隶属度函数的参数，并用此函数来实现系统离散化。

3.2　隶属度函数的构造

上海市南北高架西侧共和立交出口匝道的线圈数据被用于此次研究。数据包括上下游流量、平均速度、占有率以及是否发生交通事件，时间间隔为 5 min，共 576 条记录，其中包括事件和非事件。这里标记上游流量 $V1$，下游流量 $V2$，上游速度 $S1$，下游速度 $S2$，上游占有率 $O1$，下游占有率 $O2$。可将每个属性分为 4 类，用模糊语言来描述，可以将其表示为 Very_Low、Low、Medium、High。应用 FCM 算法数据进行聚类，然后以样本数据（分别为流量、速度、占有率）为横坐标，样本数据对各个聚类中心的隶属度为纵坐标来绘制图形，如图 2 所示。

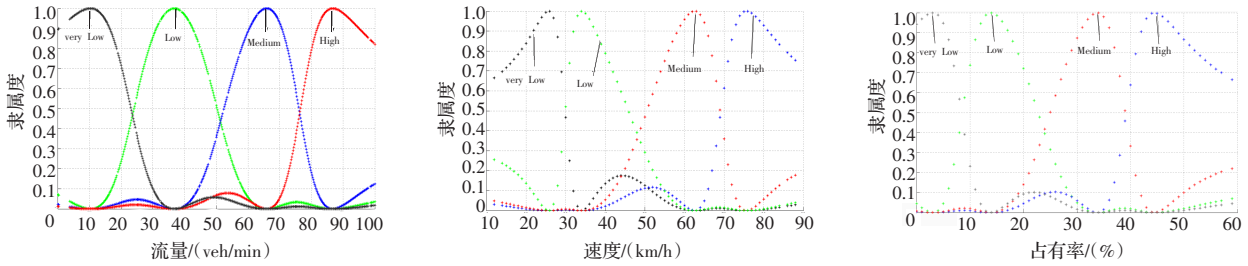

图 2　隶属度函数

表 1　拟合得到的函数及参数

		Very_Low	Low	Medium	High
流量	函数	Psigmf	Gaussmf	Gaussmf	Psigmf
	参数	[0.393 −5.2 −0.359 22.9]	[10.7 36.5]	[10.2 65.6]	[0.441 78.4 −0.369 100]
速度	函数	Psigmf	Psigmf	Psigmf	Psigmf
	参数	[0.268 12.3 −1.43 31.3]	[1.68 32.2 −0.375 45.9]	[0.237 50.4 −1.04 69.7]	[1.02 70.2 −0.224 90.9]
占有率	函数	Psigmf	Psigmf	Psigmf	Psigmf
	参数	[0.268 12.3 −1.43 31.3]	[1.14 9.27 −0.362 23.7]	[0.461 24.2 −0.969 40.6]	[0.778 39.3 −0.271 60.9]

注:函数 Gaussmf 为高斯型函数,参数按顺序[σ,c]列出,数学模型 $f(x;\sigma,c)=e^{-\frac{(x-c)^2}{2\sigma^2}}$,$x$ 是变量,σ、c 是参数。函数 Psigmf,参数按顺序[a_1 c_1 a_2 c_2]列出,是两个 Sigmoid 型函数之积:$f_1(x;a_1,c_1)*f_2(x;a_2,c_2)$,其中 Sigmoid 型隶属函数由下式给出:$f(x;a,c)=\frac{1}{1+e^{-a(x-c)}}$,$x$ 是变量,a、c 是参数。

图 2 聚类得到的隶属度函数中间的曲线非常接近高斯型,两端曲线接近 Sigmoid 型函数[11]。这并非偶然现象,经过对多组数据聚类实验发现,标准型 FCM 函数在一维上的聚类结果都具有这个特征。高斯型函数和 Sigmoid 型都具有很好的光滑性,图形没有零点,而且具有比较清晰的物理意义,是模糊系统中常用的隶属度函数之一。因此,经过对数据的分析本文采用隶属度函数 Gaussmf 和 Psigmf 来拟合隶属度函数,拟合后隶属度函数及相应的系数如表 1 所示。离散化时取变量在模糊集中的隶属度最大值对应的属性作为模糊值。通过以上模糊化处理,各节点的清晰规则集就转化为模糊规则集,即用 Very_Low、Low、Medium、High 来取代清晰规则集中的数字量。值得注意的是由于高速公路上下游的交通特性较为相似,对上下游流量、速度、占有率离散化过程中选取统一值,忽略上下游的微小差异,这样便于对决策表的进一步处理。表 2 为原始决策表和应用此方法离散后的决策表。其中 $V1$、$V2$ 代表线圈所检测到的上下游的流量,$S1$、$S2$ 代表上下游速度,$O1$、$O2$ 代表上下游占有率,D 为是否发生交通事件。条件属性值 1、2、3、4 代表相应的等级描述 Very_Low、Low、Medium、High,决策属性值 1 表示有事故发生,反之。

表 2　交通事件决策表

(a)原始决策表

U	$V1$	$V2$	$S1$	$S2$	$O1$	$O2$	D
1	92	34	21	83	56	15	1
2	67	40	76	72	13	36	0
3	83	91	13	80	61	34	0
4	45	63	37	74	50	18	1
5	12	72	15	58	21	33	0
6	36	89	42	80	47	38	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
576	60	41	78	71	13	29	0

(b)离散后的决策表

U	$V1$	$V2$	$S1$	$S2$	$O1$	$O2$	D
1	4	2	1	4	4	2	1
2	3	2	4	4	2	3	0
3	4	4	3	4	3	3	0
4	2	3	2	4	4	2	1
5	1	3	1	3	2	3	0
6	2	4	2	4	4	3	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
576	3	2	4	4	2	3	0

4　模糊粗糙集的交通事件检测

4.1　粗糙集理论的基本原理

粗糙集理论自 Z.Pawlak 于 20 世纪 80 年代提出以来已成功地应用于许多领域,它不仅具有模拟人类逻辑思维的能力,而且能有效地分析和处理不精确、不一致、不完整的信息,发现和揭示数据之间内在的规律以提取有用信息。该方法的主要优点在于它不需要预先给定某些特征或属性的数量描述和模型假定。但单纯地使用粗糙集理论不一定总能有效地描述数据不精确或不确定的实际问题,而与其他理论结合形成更加有效的方法来解决相关工程技术问题是目前研究的重点。例如:粗糙集和模糊聚类相结合[7-8],两者在处理不确定性或不精确性问题方面都推广了经典集合论,并且都可以用来描述知识的不精确性或不完全性,但着眼点不同,它们又可以相互结合,以发挥各自得优势。

定义 1　假设 R 代表论域 U 中的一种关系,当 R 描述对 U 的分类,即 U 中对象之间的等价关系时,用 $U/R=\{X_1,X_2,\cdots,X_n\}$ 表示根据关系 R、U 中的对象构成的所有等价类族,称为关于 U 的知识。$[x]_R=\{y\in U|xRy\}$ 表示关系 R 下包含元素 x 的等价类,(U,R) 称为近似空间(Approximation Space)。$X_1,X_2,\cdots,X_n$ 形成了知识 U/R 的组成颗粒,它们源于对象之间的不可分辨性。

定义 2　若 $P\subseteq R$,且 $P\neq\phi$,则 P 中的全部等价关系的交集称为 P 上的不可分辨关系,记为 $ind(P)$。$[X]_{ind(P)}=\cap[X]_R$,式中,$[X]_R$ 表示所有与 X 不可分辨的对象所组成的集合,即由 X 决定的等价类,$[X]_R$ 中的每一个对象都与 X 具有相同的属性。

定义 3　给定知识库 $K=(U,R)$,对每个子集 $X\subseteq U$ 和一个等价关系 $R\subseteq ind(K)$。集合 X 关于 R 的下近似(Lower Approximation)为:$\underline{R}(X)=\{x\in U:[x]_R\subseteq X\}$。$\underline{R}(X)$ 是由那些根据论域知识判断肯定属于 X 的对象 x 所组成的最大集合,也称为 X

的正区域(Positive Region)，记作 $POS(X)$。

关于 R 的上近似(Upper Approximation)为：$\overline{R}(X)=\{x\in U:[x]_R\cap X\neq\phi\}$，$\overline{R}(X)$是所有与 X 的交集为空的等价类$[x]_R$的并集，是由那些根据论域知识判断可能属于 X 的对象 x 组成的最小集合。

上近似、下近似的概念刻划了一个边界不清晰的集合的逼近特性。在粗糙集理论中，一个集合之所以粗糙，正是因为划分论域的知识不够充分，无法用知识的最小模块(基本等价类)准确地定义。

定义 4　对于属性子集 $P\subseteq R$，若存在 $Q=P-r$，$Q\subseteq P$，Q 是独立的(即最小子集)，使得 $ind(Q)=ind(P)$，则称 Q 为 P 的一个约简(Reduction)，记作 $red(P)$。本质上，约简 Q 是能够与 P 表达同样知识的最小等价关系集合，是 P 中的重要部分。虽然 Q 除去了部分对于 P 的知识，仍然可以取得与原有的完整知识基 P 一样的分类结果。同时，一个属性集合 P 可能有多个约简。

定义 5　属性集合 P 的所有约简的交集定义为 P 的核(Core)，记作 $core(P)$。显然，属性集合的约简和核的关系如下：$core(P)=\cap red(P)$。

$core(P)$含有 P 的全部约简中共同的等价关系，是属性集合 P 中不可缺少的重要属性集。通过属性约简，可以将决策表中对决策不必要的属性省略，从而实现决策表的简化，有利于从决策表中分析发现对决策分类起作用的属性。

定义 6　对于信息系统 $S=(U,A,V,f)$，$A=C\cup D$，C 称为条件属性，D 称为决策属性。经过属性约简后，如果去除任一属性值 $v\in V$，决策规则发生冲突(不一致或不相容)，则该属性值 v 是不可省略的。否则，该属性值可省略。

4.2　推理规则的建立

例中数据经离散化和模糊化的决策表，如表 2(b)有 576 个对象，如对决策表不进行属性约简，那么模糊推理计算量将非常巨大，这里采用了粗糙集理论来建立模糊推理规则集[12-14]，简化决策表即属性约简和属性值约简，约简前后不改变决策表的决策属性。首先对决策表进行预处理，合并表中的相同的元素，元素 2 和 576 的条件和决策属性相同，将其合并；针对冲突的元素采取大概率的原则，即有相同的条件属性值不同的决策属性值，则选取具有数量较多规则的决策属性作为统一的决策属性值；最后去除 6 个属性中的任一属性，若去除后属性表仍为一致性，则该条件属性可以忽略，反之不可以忽略。经计算约简后的属性为 $S1$、$S2$、$O1$、$O2$，即上下游速度和占有率可以决定所处的交通状态。另外，为了更加简洁，需对每一决策规则进行属性值简化，针对每条决策规则计算其核值，去除规则中不必要的属性值。经粗糙集约简后的决策表如表 3 所示。

表 3　简化的决策信息表

U	$S1$	$S2$	$O1$	$O2$	D
1	1	4	*	*	1
2	4	4	*	*	0
3	3	*	*	*	0
4	2	4	4	*	1
5	1	*	2	*	0
⋮	⋮	⋮	⋮	⋮	⋮
61	2	4	*	3	1

从表 3 能较容易的得到其推理规则：

规则 1　If 上游速度为 Very_Low and 下游速度为 High then 有事件发生；

规则 2　If 上游速度为 High and 下游速度为 High then 无事件发生；

规则 3　If 上游速度为 Medium then 无事件发生；

……

规则 61　If 上游速度为 Low and 下游速度为 High and 下游占有率为 Medium then 有事件发生。

4.3　交通事件的模糊推理

模糊推理运算采用最普遍的 Max-Min 模糊推理，即应用 Max-Min 模糊推理，获得当前交通状态下的 61 条规则中每条规则的条件属性对应的隶属度最小值，并将这个最小值分配给每条规则，该最小值为规则所对应交通模式模糊推理输出值。再找到模糊推理输出值中最大的规则，该规则为当前所识别的交通事件模式。

例某时刻检测到此路段实时交通流数据为上游速度 19 km/h，下游速度 78 km/h，上游占有率为 43%，下游占有率为 3%。按照表 1 的隶属度函数模糊化后的上游速度(0.821，0，0，0)，下游速度(0，0，0，0.903)，上游占有率(0，0，0.1，0.851)，下游占有率(0.854，0，0，0)，那么采用 Max-Min 模糊推理根据表 3 的 61 条推理规则，可得此时该路段上可能有交通事件发生。

5　算法验证及分析

一般用于评价交通事件检测算法的指标有检测率(DR)、误报率(FAR)和平均检测时间(MTD)[4]。当选择算法时，通常要在这些性能量度之间进行权衡，要达到高的检测率必须设置高的误报率。为减少事件响应时间，必须接受相对较高的误报率，并接受附加的计算成本，可见一些指标在本质上是不相容的。

本文在内存为 1 G，CPU 为 Pentium3.06 G，操作系统为 Windows 2000 Professional 的硬件环境，利用 Matlab 实现了上述算法，利用上述实测数据，对算法进行验证，并与几种传统的算法进行比较，结果如表 4。从对比实验看出，利用本文所提出的识别方法检测率高于其他三种算法，误报率低于除加利福尼亚算法外的三种方法，而平均检测时间低于其它三种方法，在整体性能指标上优于其它方法。

表 4　算法结果比较

算法	DR/(%)	FAR/(%)	MTR/s
加利福尼亚算法	49	0.57	210
双截面 McMaster 算法	37	2.07	255
Minnesota	76	3.05	229
本文所提出的模型	79	1.64	200

6　结论

本文提出了基于 FCM 和模糊粗糙集理论的交通事件检测的新模型，系统通过聚类算法实现了对输入空间的划分和隶属度函数类型及参数的确定，利用模糊集和粗糙集理论以条件属性的模糊性和不可分辨关系为基础，进行了属性约简，并提取了交通事件的规则，组成了推理的规则库，简化了推理规则数目，从而对交通事件进行推理。实验结果表明，本文提出的交通事件自动检测模型是正确有效的，且总体性能优于一般的事件自动检测(AID)方法，从而为 AID 提供了一种更科学有效的方法。

(下转 21 页)