支撑交通管理综合信息平台的信息挖掘模型

周海淞[1] 朱 茵[2] 陆化普[3]

1．杭州市公安交通警察支队，科研所，杭州 310014

2．中国人民公安大学，交通管理工程系，北京 102614

3．清华大学，土木工程系交通所，北京 100084

摘 要：本文分别从结构化信息源和非结构化信息源两个角度探讨支撑交通管理综合信息平台的信息挖掘模型，提出利用粗糙集理论的基本原理，对于其属性约简进行改进。在保证原有属性不变的条件下，削减了冗余信息，并可实现基于数据库存储数据的动态数据挖掘模型，避免了将大量数据从数据库中基于主题倒入数据仓库的传统数据挖掘模式。通过上述算法，不仅提高了信息挖掘的效率，而且可以实时在线信息挖掘，相对于传统的挖掘模式，本文提出的方法更适用于当前基于网络的交通管理综合信息平台的实际应用。

关键词：信息挖掘；交通管理；综合信息平台

中图法分类号：U491 文献标识码：A 文章编号：1672-4747(2005)02-0027-08

An Information-Mining Model Supporting Traffic Management Integrated Information Platform

ZHOU Hai-song[1] ZHU Yin[2] LU Hua-pu[3]

1. Science Research Institute,
Traffic Police Detachment of Hangzhou,
Hangzhou 310014, China

2. Department of Traffic Management Engineering,
Chinese People's Public Security University,
Beijing 102614, China

3. Institute of Transportation Engineering,

收稿日期：2005-01-05.

作者简介：周海淞（1963-），男，浙江省人，杭州市公安局公安交通警察支队科研所，所长，研究方向：智能交通系统工程。

Civil Engineering Department,
Tsinghua University, Beijing 100084, China

Abstracts：An information-mining model supporting the traffic management integrated information platform was discussed from two aspects of structured or unstructured information sources. By using the relative rough set theory, the reduced property was improved in reducing the redundant information under the condition of guarantee of the model original functions. Then, the modified information-mining model can directly obtain information from the database otherwise data ware, and processes the real-time and on-line information mining with high effectiveness.

Key words：Information mining；traffic management；integrated information platform

0　引　言

随着信息采集、信息存储设备的飞速发展，城市智能交通管理综合信息平台已经有了大量的相关信息做支撑。然而，目前更多的是在数据处理层面上只做简单的处理，缺乏相应的手段来挖掘数据背后隐藏的知识，无法发现并利用数据中存在的关系和规则，从而根据现有的数据预测未来的发展趋势，并进一步辅助决策者更好地利用有效数据进行科学决策。在这种现状下，需要一种新的技术方法，即信息挖掘来充分挖掘这些有效信息，为基于综合信息平台的上层应用奠定基础。

对于信息挖掘，本文采用如下含义：

信息挖掘是指从各种各样的信息源（包括结构化的和非结构化的信息源）中，抽取先前未知的、完整的信息，来做关键的业务决策。因此，本文分别从结构化信息源和非结构化信息源两个角度探讨支撑交通管理综合信息平台的信息挖掘模型[1]。

1　信息挖掘模型在综合信息平台中应用的优势

综合信息平台的建设在智能交通系统的研究与开发建设中得到了越来越多的关注，但是，由于平台信息的来源既包括既有的相关管理信息系统，例如：机动车驾驶员管理信息系统、机动车管理信息系统、事故管理信息系统、违法关系信息系统等等，同时也包括通过各种交通信息传感器采集的交通流量信息、平均运行速度信息等等，还包括基于网络的办公管理系统的文本类、时间序列、Web 信息等非结构化或半结构化信息。同时，当前的综合信息管理平台为了提供更好的可视化显示效果，更多的采用了基于 WebGIS 的交通状况显示功能，因此，又增加了地理信息相关的空间信息。尽管平台集成了大量的综合信息，但是，究竟如何将上述信息之间的关系充分关联起来，最大限度的挖掘出其中的内在知识与规律，却是当前未能全面解决的问题，因此，有必要通过有效的信息挖掘模型，其中包括结构化数据挖掘模型、复杂类型数据挖掘模型的综合利用，将上述综合信息存在的内在知识与规律挖掘出来。

2　结构化数据挖掘模型

大量的数据挖掘算法是基于数据仓库的基础上形成的，而在城市智能交通管理综合信息平台的研究与开发中，往往很多数据是存储在综合信息数据库中，而未倒入数据仓库中，从这一意义上而言，本系统对于动态数据挖掘的需要更强于基于数据仓库的挖掘算法，基于此，本文提出利用粗糙集（Rough Set，RS）理论的基本原理，对于其属性约简进行改进，并可实现基于数据库存储数据的动态数据挖掘模型。

（1）RS 相关基础理论

RS理论是一种刻画不完整性和不确定性的数学工具，能有效地分析和处理不精确、不一致、不完整等各种不完备信息，并从中发现隐含的知识，揭示潜在的规律[2][3]。

但是，由于在RS中的规则约简中，只能保证确定规则在约简前后不被改变，就必然会带来很大的局限性。因为，在实际系统的应用中不确定规则是较为普遍存在的，所以，有必要通过相应算法的改变，保证确定规则与不确定规则在约简前后均不会被改变，改进后的约简算法，同时保留确定规则和不确定规则的可信度，更加符合实际需求。另外，在实际应用中，为减少动态概念属性约简过程的运算时间，一般可先在离线状态下进行底层属性约简，再根据用户需求实现动态的在线属性约简。

设$S=<U, R, V, f>$为一知识表示系统，其中：U为论域；R为属性集合；V为属性集合，$V=U_{r\in R}V_r$，V_r表示属性$r\in R$的属性值范围；$f:U\times R\to V$是一个信息函数，它指定U中每一个对象x的属性值。

设D_1，D_2，…，D_n为n个论域，又设$S(D_1)$，$S(D_2)$，…，$S(D_n)$分别为D_1，D_2，…，D_n的幂集的子集。做笛卡尔积：

$$S(D_1)\times S(D_2)\times\cdots\times S(D_n)$$

设在如上的一个知识表示系统中，若有$V=(x_1, x_2, \cdots, x_n)$，式中：$x_i\in S(D_i)$，$i=(1, 2, \cdots, n)$，则称$r$的元组的属性值为多值，称知识表示系统$S$为多值关系的知识表示系统，若$S(D_1)\times S(D_2)\times\cdots\times S(D_n)$分别为$D_1$，$D_2$，…，$D_n$的幂集的模糊子集，则称知识表示系统$S$为具有模糊多值关系的知识表示系统。

设$(L, \leq)$为半序集，$L_0\in L$，若对于$\forall\alpha, \beta\in L_0$有数$D\left(\frac{\beta}{\alpha}\right)$对应，且满足：$0\leq D\left(\frac{\beta}{\alpha}\right)\leq 1$。

当$\alpha\leq\beta$时　　$D\left(\frac{\beta}{\alpha}\right)=1$

当$\alpha\leq\beta\leq\gamma$时　　$D\left(\frac{\alpha}{\gamma}\right)\leq D\left(\frac{\alpha}{\beta}\right)$

称D为L上的包含度。

显然，对于集合A，$B\in U$，若取“$\leq$”=“$\subseteq$”，则$\frac{|A\cap B|}{|B|}$为U上的一个包含度。

设X为对象集合，H为属性集合，$P\in F(H)$称为一个命题。如果映射$g:F(H)\to F(X)$，满足条件：$P\subset Q(P, Q\in F(H))\Rightarrow g(Q)\subset g(P)$，称$g$为外延映射。$D$为$F(H)$上包含度，称$T(P)=P(g(P))$为$P$的真值，$T(P\to Q)=D\left(\frac{g(Q)}{g(P)}\right)$为$P\to Q$的蕴含度。

（2）问题分析及分类规则动态计算方法

设$S=<U, R, V, f>$为一知识表示系统，各元组的定义如上。设$P_k(1\leq k\leq n)$为决策属性D上的决策区间，分别构造属性集R上属性值x_i、集合R_i、决策区间D_k的对应的概念$g_R(x_i)$、$g_R(R_i)$、$g(D_k)$，有$T_x(g_R(x_i))=1$，$x=x_i$，其中：$T_x(g_R(x_i))$为x在概念$g_R(x_i)$上的真值，构建偏序关系使得$(g(\bullet)\leq)$为一半序集，令D为$g(\bullet)$上的包含度。由蕴含度定义有：

$$T_R(x_i\to D_k)=D\left(\frac{g(D_k)}{g_R(x_i)}\right)\qquad T_R(R_j\to D_k)=D\left(\frac{g(D_k)}{g_R(R_j)}\right)$$

根据模糊推理原理，记x有概念$g_R(x_i)$使：

$$T_x(D_k)=\vee(T_x(g_R(x_i))\wedge T(x_i\to D_k))$$

$$=\vee\left(T_x(g_R(x_i))\wedge D\left(\frac{g(D_k)}{g_R(x_j)}\right)\right)$$

简记　　$D\left(\frac{g(D_k)}{g_R(x_i)}\right)$为$D_k(g_R(x_i))$

当$1\leq k\leq n$时，有$g_R(x_i)$关于决策属性D上决策区间D_k的真值向量：

$$\{D_1(g_R(x_i)), D_2(g_R(x_i)), \cdots, D_n(g_R(x_i))\}$$

知识表示系统S转化为一个具有多值关系的知识表示系统S'。

同理，$g_R(R_i)$有概念关于决策属性D上决策区间D_k的真值向量：

$$\{D_1(g_R(R_i)), D_2(g_R(R_i)), \cdots, D_n(g_R(R_i))\}$$

对于决策区间R有概念$g_R(R_i)$使得：

$$T_{R_j}(D_k)=\vee(T_x(g_R(R_j))\wedge T(R_j\to D_k))$$

$$=\vee\left(T_R(g_R(R_j))\wedge D\left(\frac{g(D_k)}{g_R(R_j)}\right)\right)$$

取$T_x(g_R(x_i)) = \mathrm{sim}(x, x_j)$，其中，sim为一定义在属性集$R$上的相似度，定义如下偏序关系：

$$g(a) \leq g(b) \Leftrightarrow \{a\} \subseteq \{b\}$$

式中，$\mu_{\{a\}}(x) = \begin{cases} 1 & \mathrm{sim}(x,a) > 0 \\ 0 & \text{else} \end{cases}$

显然，满足自反、传递和反对称性。

由相似度性质有$T_x(g_R(x_i))=1$，$x = x_i$。

$$T_R(R_j \to D_k) = D\left(\frac{g(D_k)}{g_R(R_j)}\right) = \vee(T_{x_i}(g(D_k)) \wedge T_{x_i}(g_R(R_j)))$$

$$(x_i \in U) = \vee D\left(\frac{g(D_k)}{g_R(R_j)}\right) \wedge T_{x_i}(g_R(R_j)) \quad (x_i \in g_R(R_j))$$

为简化计算可取

$$T_{x_i}(g_R(R_j)) \approx \mu_{R_j}(x)$$

当决策属性D为离散变量$(g(D_k)=D_k)$，若有

$g(R_i)=R_j$，$(R_i \cap R_j = \phi, i \neq j)$，$T_x(g_R(R_j)) = \begin{cases} 1 & x_j \in R_j \\ 0 & x_j \notin R_j \end{cases}$ 时：

$$T_R(R_j \to D_k) = D\left(\frac{g(D_k)}{g_R(R_j)}\right) = \vee(T_{x_i}(g(D_k)) \wedge T_{x_i}(g_R(R_j)))$$

$$(x_i \in U) \approx T_{x_i}(g(D_k)) \quad (x_i \in R_j) = \vee\left(D\left(\frac{g(D_k)}{g_R(x_i)}\right)\right) \quad (x_i \in g_j)$$

■ 若 $T_x(g_R(x_i)) = \begin{cases} 1 & x = x_i \\ 0 & x \neq x_i \end{cases}$，则当决策属性$D$为离散变量$(g(D_k) = D_k)$时，$D\left(\frac{g(D_k)}{g_R(x_i)}\right) = \begin{cases} 1 & D(x_i) \in D_k \\ 0 & \text{else} \end{cases}$，上述方法退化为一般离散化方法。

■ $T_x(g_R(x_i)) = \begin{cases} 1 & x = x_i \\ 0 & x \neq x_i \end{cases}$，则

$$D\left(\frac{g(D_k)}{g_R(x_i)}\right) = \begin{cases} T_{x_i}(g(D_k)) & D(x_i) \in D_k \\ 0 & \text{else} \end{cases}$$

当决策属性D为离散变量$(g(D_k)= D_k)$时

$$D\left(\frac{g(D_k)}{g_R(x_i)}\right) = \begin{cases} 1 & D(x_i) \in D_k \\ 0 & else \end{cases}$$

若有$T_x(g_R(R_j)) \in [0, 1]$，即退化为一般模糊概念方法。

■ 若 $T_x(g_R(x_i)) = \begin{cases} 1 & x = x_i \\ 0 & x \neq x_i \end{cases}$，

$$D\left(\frac{g(D_k)}{g_R(x_i)}\right) = \begin{cases} T_{x_i}(g(D_k) & D(x_i) \in D_k \\ 0 & \text{else} \end{cases}$$

若$T_x(g_R(R_j)) \in [0, 1]$，即退化为模糊概念和模糊关系。

从以上运算可以看出，由于$T_{x_i}(g(D_k))$为该分类的系统特性，$T_R(R_j \to D_k)$只与R_j的选取有关，因此，可以进行动态分类的规则学习。

当采用不同的包含函数计算得到的$T_R(R_j \to D_k)$有所不同，反映了不同的规则理解和推理方法需要的差异，同时对于不同的$T_R(R_j \to D_k)$在R_j上信息的冗余进行不同的平均和消减。

若取

$$D\left(\frac{g(D_k)}{g_R(R_j)}\right) = \frac{1}{m}\sum_{i=1}^{m}(g(D_k)(x_i) \wedge g_R(R_j)(x_i))$$

当R_j为经典离散化区间时有：

$$D\left(\frac{g(D_k)}{g_R(R_j)}\right) = \frac{\sum_{i=1}^{m} D_k(g_R(x_i))}{m} \quad x_i \in R_j$$

因此，得到R_j的分类规则向量：

$$\{T_R(R_j \to D_k)\} \quad 1 \leq k \leq m$$

根据广义模糊推理原理，利用R_j的分类规则向量$\{T_R(R_j \to D_k)\}$，$1 \leq k \leq m$可进行如下推理：

■ 肯定一个结论的陈述；

■ 肯定一个结论的所需的可信度；

■ 认定推理的可信度尽量大的结论。

（3）属性约简

属性约简一般是指原有分类规则和规则强度不变的前提下，对条件属性中的冗余属性进行约简的过程。为描述方便，将属性约简划分为两种约简形式：底层属性约简和动态概念属性约简，并在粗糙集理论中属性约简的框架下给出了其相应的定义和部分性质。

■ 底层属性约简

定义 1　设S'为一模糊多值知识表示系统（定义同前），若存在条件属性A，使得对于任意的$x \in U$和$D_k(1 \leq k \leq m)$，有$T_C(x \to D_k) = T_{C-A}(x \to D_k)$，则称属性$A$是条件属性集$C$中依属性值可省略的，否则称属性$A$是条件属性集$C$中依属性值不可省略的。

定义 2　设S'为一模糊多值知识表示系统（定义同前），若条件属性集C中所有的条件属性都是依属性

值不可省略的，则称条件属性集C是依属性值独立的，否则，称条件属性集C是属性值依赖的。设存在$Q \subseteq C$，若Q是独立的，且对于任意$x \in U$和$D_k(1 \leq k \leq m)$，有$T_C(x \rightarrow D_k)=T_Q(x \rightarrow D_k)$，则称$Q$是$C$依属性值的一个约简。

■ 动态概念属性约简

底层属性约简是在数据底层进行的一种属性约简形式，其中保留了大量的数据原始信息。然而，在实践过程中，这样的信息不但不便于表示和理解，而且也不符合实践过程的需要。对其中的信息进行合并，投影到用户定义的概念上往往是有必要的。

定义 3　设R是U上的一个等价关系，U/R表示R的所有等价类构成的集合，表示包含元素的等价类，$[x]_R(t)$表示元素的真值，其中$[x]_R(t)$表示$[x]_R(t)$包含元素 $x \in U$ 的第i个等价类。设$U/R = \{x_i\}$，$(1 \leq i \leq n)$，$X_i \subseteq U$，$X_i \neq \phi$，$\bigcup_{i=1}^{n} X_i = U$ 称x_i为属性R上的一个概念。若有$X_i \cap X_j = \phi$，$(i \neq j)$，$T(x \mid X_i) = \{0, 1\}$则称X_i为属性R上的一个经典概念，否则，称X_i为属性R上的一个模糊概念。设$R_k(X_i)$为属性$R_k \in R$上的一个概念，称$\bigcap_{k=1}^{|R|} R_k(X_i)$为属性集 R 上的一个概念。

定义 4　设S'为一模糊多值知识表示系统（定义同前），A_i为属性 $A \in R$ 上的概念，R_j为属性集R上的概念。若有$T_R(R_j \rightarrow D_k) = T_{(R-A)}((R_j - A_i) \rightarrow D_k$，$(1 \leq k \leq n)$，则称论域$U$上概念$A_i$相对于概念$R_j$是可省略的，反之，若论域$U$上概念$A_i$相对于概念$R_j$是不可省略的。若论域$U$上概念$A_i$相对于属性集$R$上的任一概念均是可省略的，则称论域$U$上概念$A_i$是依概念族$\{R_j\}$可省略的，反之称论域$U$上概念$A_i$是依概念族$\{R_j\}$不可省略的；若属性$A$任一概念在论域$U$上均是依概念族$\{R_j\}$可省略的，则称在论域$U$上属性$A$是依概念族$\{R_j\}$可省略的，反之称在论域$U$上属性$A$是依概念族$\{R_j\}$不可省略的。

定义 5　设A_i为属性 $A \in R$ 上的概念，若存在属性B，使得在论域U上概念A_i相对于概念族$\{A_i \cap B_j\}$是可省略的，则称概念A_i是依赖于概念族$\{B_j\}$，简记为$A_i \in \{B_j\}$；若属性A任一概念在论域U均是依赖于概念族$\{B_j\}$，则称在论域U上概念族$\{A_i\}$依赖于概念族$\{B_j\}$，简记为$\{A_i\} \subseteq \{B_j\}$。

在实际应用中，为减少动态概念属性约简过程的运算时间，一般可先在离线状态下进行底层属性约简，再根据用户要求动态的在线进行动态概念属性约简。

3　复杂类型数据挖掘

■ 复杂类型数据挖掘是相对于结构化数据挖掘提出的，即网络信息挖掘。前文已经论述，对于网络办公管理信息的挖掘，更多的是集中在文本内容的挖掘，通过挖掘模型的实施，可以实现网络办公管理信息的层次性组织，同时，可以结合对用户访问日志记录信息的挖掘，把握用户的兴趣，从而有助于开展个人信息定制服务等等，既方便了网络系统的管理，同时也方便了用户的使用。

网络信息挖掘主要由信息采集、特征提取和特征匹配 3 部分构成。

■ 源信息采集

WWW 是以超文本的形式存储信息并提供信息服务的，在 WWW 上进行源信息采集，需要通过 Robot 程序实现。Robot 是一个能沿着 Web 页面中的超链接进行自动漫游的程序，并且能够通过 HTTP 等标准协议下载所漫游到的页面。WWW 是一个网状结构的信息空间，我们可将其作为一个有向图处理：将页面作为图中的节点，页面中的超链接作为图中的有向边。因此，我们可以使用有向图遍历算法（深度优先算法和广度优先算法）对其进行遍历。源信息采集是进行网络信息挖掘的重要环节。为了提高挖掘的效率，在源文档采集阶段就应对信息源进行一定的过滤。

目标表示与特征匹配目标表示是指以一定的特征项（如词条或描述）来代表目标信息，在信息挖掘时用这些特征项评价未知文档与用户目标的相关程度，目标表示的构造过程就是挖掘模型的构造过程。目标表示模型有多种，常用的有布尔逻辑型、向量空间型、概率型等。近年来应用较多且效果较好的目标表示法是向量空间模型(Vector Space Model，VSM)法。

在VSM中，将文本文档看作为是由一组词条(T_i，T_2，…，T_n)构成，对于每一词条T_i，都根据其在文档中的重要程度赋以一定的权值W_i。我们可以将其看成一个n维坐标系，W_1，W_2，…，W_n为对应的坐标值，因此，每一篇文档都可映射为由一组词条矢量张成的向量空间中的一个点。对于所有用户目标或未知文档都可用词条特征矢量(T_1，W_1，T_2，W_2，…，T_n，W_n)表示，从而将文档信息的匹配问题转化为向量空间中的向量匹配问题处理。假设用户目标为U，未知文档为V，两者的相似程度可用向量之间的夹角来度量，夹角越小说明相似度越高，相似度计算公式如下：

$$\mathrm{Sim}(V,U)=\cos(V,U)=\frac{\sum_{k=1}^{n}W_{vk}\cdot W_{uk}}{\sqrt{\sum_{k=1}^{n}W_{vk}^2}\cdot\sqrt{\sum_{k=1}^{2}W_{uk}^2}}$$

■ 特征提取

目标表示中词条 T 及其权值的选取称为特征提取，特征提取是挖掘目标共性与规则的提取过程，其采用策略的优劣将直接影响到挖掘工具的效果。词、词组和短语是组成文档的基本元素，并且，在不同内容的文档中，各词条出现频率有一定的规律性，因此，可根据词条的频率特性进行目标特征提取。

构造词条权值评价函数：

$$W_{ik}=\frac{tf_{ik}\log\left[\frac{N}{n_k}+0.01\right]}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2\cdot\log^2\left[\frac{N}{n_k}+0.01\right]}}$$

式中：tf_{ik}表示词条T_k在文档D_i中的出现频数；N表示全部样本文档总数；n_k表示词条T_k的文档频数。

网络信息挖掘所处理的对象一般都是 HTML 文档，HTML 文档中存在很多标记信息，这些标记信息往往对文档的内容有很高的概括性，因此，可利用这些标记信息提高特征提取精度。在特征提取时，可设置 CofTitle，CofLinkText，CofH1，CofH2 等一系列针对 HTML 文档中的{<Title>，</Title>}、{<A>，</A>}、{<Hi>，</Hi>}等域文本的加权系数，对出现在不同域的词条赋以不同的频率加权系数。

■ 特征提取

文本信息的预处理

在对文档进行特征提取前，需要先进行文本信息的预处理，这主要包括英文文档的 Stemming 处理和中文文档的词条切分。从英文单词的多种形式中提取出其基本词干的过程，被称作“ Stemming ”。英文单词在具体使用时，可以有现在时、过去时等多种形式，如“ walk ”，“ walked ”，“ walker ”，“ walking ”，还有的单词有名词、形容词、副词等多种形式，如“ use ”，“ useful ”，“ usefulness ”，“ usefully ”等，但它们的词干是相同的，因此，在进行词频统计时应该作为相同的词处理。实现 Stemming 一般的方法是建立单词前缀、后缀表和特殊形式表，用匹配方式实现。

中文信息的处理与英文不同，句子中各词语间没有固有的分隔符(空格)，因此，在进行词频统计等处理前，先要对中文文档进行词条切分处理，中文文本的分词就是在中文文本的各词条间加入分隔符，将中文文本的连续字流形式转化为离散的词流形式。中文文本的分词方法有很多种，各种方法适用的情况也不同，网络信息挖掘对分词处理要求有较高的实时性，但对分词的准确度不太敏感，容许一定的分词错误率，因此，可以采用较为简单的基于词典的正向匹配、逐词遍历分词方法。比较简单有效的分词方法，是基于词表的机器分词法，这需要建立大型的切分词库。在进行词频统计时，还应考虑到自然语言的多样性，建立并使用相应的同义词词典、蕴含词词典等辅助词典，以提高挖掘的准确度。

非文本信息处理

在 WWW 中，有很多图像信息和以 PDF，PS 等格式存储的文档，如果采用图像处理和 OCR 的方法对其进行内容分析和特征提取，将会使系统变得很十分庞大和低效。考虑到 WWW 中的非文本信息一般都是采用“ 链接-文件 ”对的形式呈现给用户的，每个文件都有一段链接文本(关于链接的描述文本，如出现在<A>，</A>标记对间的文字)与其对应，而这些链文本往往都是对所链接的非文本对象的高度概括描述，所以，可以采用非文本文件的链文本对其进行特征提取，从而将非文本信息转化为文本信息进行处理。

4　算例分析

对于本文提出的模型中，在综合信息平台中最为常用的是结构化数据挖掘模型，而其中更为核心的是基于粗糙集的方法。对于粗糙集体理论，最大的优势集中在属性约简。对此，本文提出一个算例，来进一步说明其计算流程。

设关系数据库中有一对应机动车驾驶员违法行为的决策表，$U=\{x_1, x_2, x_3, \cdots, x_9\}$，论域中的的9个元素分别代表关系数据库中的9条记录，其对应的条件属性为$R=\{r_1, r_2, r_3, r_4\}$，即代表着对应机动车驾驶员的4种违法行为，并且，上述元素为彼此独立的，上述行为对应的决策属性集为$D=\{d\}$，并且，假设当前的9条记录出现的频率是均等的，即为1/9，上述具体信息如表1（表中的ID为关系数据库中机动车驾驶员的主键ID）所示：

表1　关系数据库原始信息

Tab.1　Original Information of RDS

ID	r_1	r_2	r_3	r_4	d	频 率
x_1	1	0	2	1	0	1/9
x_2	0	0	1	2	1	1/9
x_3	2	0	2	1	0	1/9
x_4	0	0	2	2	0	1/9
x_5	1	1	2	1	0	1/9
x_6	1	0	2	1	0	1/9
x_7	1	0	2	1	0	1/9
x_8	2	0	2	1	0	1/9
x_9	1	0	2	1	0	1/9

从表1中可见，记录$\{x_1, x_6, x_7, x_9\}$和记录$\{x_3, x_8\}$分别具有相同的条件属性和决策属性，因此，可将其合并，合并后的表1转换为表2所示：

表2　经过第一次属性约简后的数据库信息

Tab.2　The first simplified database Information

ID	r_1	r_2	r_3	r_4	d	频率
x_1	1	0	2	1	0	4/9
x_2	0	0	1	2	1	1/9
x_3	2	0	2	1	0	2/9
x_4	0	0	2	2	0	1/9
x_5	1	1	2	1	0	1/9

根据表2对条件属性进行约简，由于：

$$T_{\{r_1,r_2,r_3,r_4\}}(x\to e)=T_{\{r_1,r_2,r_3\}}(x\to e)=\{x_1,x_2,x_3,x_4,x_5\}$$

所以，条件属性r_4可消去，则表2转换为表3。

表3　经过第二次属性约简后的数据库信息

Tab.3　The second simplified database Information

ID	r_1	r_2	r_3	d	频率
x_1	1	0	2	0	4/9
x_2	0	0	1	1	1/9
x_3	2	0	2	0	2/9
x_4	0	0	2	0	1/9
x_5	1	1	2	0	1/9

又由于：

$$T_{\{r_1,,r_3\}}(x\to e)=T_{\{r_1,r_2,r_3\}}(x\to e)=\{x_1,x_2,x_3,x_4,x_5\}$$

所以，条件属性r_2可消去，则表3转换为表4：

表4　经过第三次属性约简后的数据库信息

Tab.4　The third simplified database Information

ID	r_1	r_3	d	频率
x_1	1	2	0	4/9
x_2	0	1	1	1/9
x_3	2	2	0	2/9
x_4	0	2	0	1/9
x_5	1	2	0	1/9

由于记录x_1与记录x_5相同，所以，将表中相同记录进一步合并得到表5。

表5　经过四次约简后的最终数据库信息

Tab.5　The final simplified database Information

ID	r_1	r_3	d	频率
x_1	1	2	0	5/9
x_2	0	1	1	1/9
x_3	2	2	0	2/9
x_4	0	2	0	1/9

表5中$T_{\{r_1,,r_3\}}(x\to e)\neq T_{\{r_1\}}(x\to e)\neq T_{\{r_2\}}(x\to e)$，所以，剩余的条件属性不能再继续约简，因此，表5为最终

下转第38页

4　结束语

本文提出了军事运输中存在的路段弧容量上限为区间数时的不确定性网络优化问题；建立了保守最大流、乐观最大流、最小风险代价乐观最大流以及最小风险代价流等关于网络优化问题的新定义及其数学模型；针对目标为非线性函数的优化问题，设计了特殊的算法，即通过可调圈求最大流问题和给定流配流问题的多重解，从而获得最小风险代价乐观最大流和最小风险代价流；最后，给出的算例证明了算法的有效性。

参考文献

[1] 谢金星. 刑文训. 网络优化[M]. 北京：清华大学出版社，2000.

[2] Bondy J. A., Murty USA. Graph theory with applications [M]. New York，American Elserver：1976.

[3] 林景荣. 最小费用流问题的多重最优解[J]. 海南大学学报，1994；12（2）：103-107.

[4] 韩明亮. 求解最小费用最大流问题的一种方法[J]. 中国民航学院学报，2000；18（1）：49-53.

上接第 33 页

约简后的决策表。

通过上述属性约简过程，可以在保证记录的原有分类规则和规则强度不变的前提下，对条件属性中的冗余属性进行合理的约简，得到最终更为有效的决策表，特别是对于大量属性存在的条件下，该方法的优势将更为突出。

5　结束语

本文针对支撑城市智能交通管理综合信息平台的信息挖掘模型进行了研究与探讨，根据交通管理综合信息平台的实际需求，分别针对结构化数据和复杂类型数据的挖掘模型进行了阐述，并分别针对上述数据需求阐述了相关挖掘模型。本文提出的利用粗糙集理论的基本原理，对于其属性约简进行改进，在保证原有属性不变的条件下，削减了冗余信息，并可实现基于数据库存储数据的动态数据挖掘模型，避免了将大量数据从数据库中基于主题倒入数据仓库的传统数据挖掘模式。由于本文提出的算法是基于数据库存储信息进行挖掘的，因此，其挖掘过程可以实时在线信息挖掘，而传统的基于数据仓库的挖掘模式，仅仅是对历史数据的挖掘，本文提出的方法更适用于当前对于实时性和挖掘信息的效率的要求都很高的交通管理综合信息平台的实际需求。

参考文献

[1] 朱　茵. 智能交通管理信息系统相关理论模型的研究（博士后出站研究报告）[R]. 北京：清华大学，2004.

[2] 苗夺谦. Rough Set 理论中连续属性的离散化方法[J]. 自动化学报，2001；27（3）：296-302.

[3] 赵　军，王国胤，吴中福，李　华. 基于粗集理论的数据离散化新算法 [J]. 重庆大学学报（自然科学版），2002；25（3）：18–21.

[4] 邹　涛，王继成，张福炎. 基于 WWW 的资料搜集系统的设计与实现 [J]. 情报学报，1999；18（3）：195-201.