可拓数据挖掘的概念与理论

黄金才　陈文伟
(国防科技大学信息系统与管理学院,长沙 410073)
E-mail:huangjincai@nudt.edu.cn

摘　要　论文从数据挖掘概念和理论拓宽到可拓数据挖掘概念和理论,证明了两个可拓数据挖掘定理,并通过实例说明可拓数据挖掘是:在数据挖掘中获取的知识的基础上,通过可拓变换,获取可拓变换规则知识(变化知识)。

关键词　数据挖掘　可拓变换　可拓数据挖掘

文章编号 1002-8331-(2006)14-0007-02　**文献标识码** A　**中图分类号** TP311

The Conception and Theory of Extension Data Mining

Huang Jincai　Chen Wenwei
(School of Information System and Management, National University of Defense Technology, Changsha 410073)

Abstract: This paper extends the concept and theory of data mining to extension data mining, and proves two theorems of extension data mining.Explain the extension data mining through example as: based on the knowledge that obtained from data mining, obtain the extension transformation rule knowledge(changing knowledge) by means of extension transformation.

Keywords: data mining, extension transform, extension data mining

1　数据挖掘概念与理论

1.1　数据挖掘概念

数据挖掘可以概括的理解为:从数据库的数据中获取知识。具体的表示为:对数据库中元组之间进行分析,获取聚类知识(按距离标准),或分类知识(是否覆盖),或者对属性之间进行相关分析,获取关联规则或发现变量(属性)间满足的数学公式。关系数据库可以看成是大量基元的相同特征(属性)取不同特征值(元组)的集合。数据挖掘是对已有的元组和属性数据,通过聚类、分类、相关分析等获取知识(主要是规则知识)。

1.2　数据挖掘理论

对于只具有正例和反例两类数据的数据库,通过数据挖掘的分类算法,如归纳学习(示例学习、遗传算法等)方法,可以得到正例与反例的规则:

(1)正例规则

$$\wedge a_i \rightarrow P \tag{1}$$

其含义是:当某元组的属性 $a_i(i=1,2\cdots)$ 同时成立(与∧)时,则该元组属于正例类 P。

(2)反例规则

$$\wedge b_j \rightarrow N \tag{2}$$

其含义是:当某元组的属性 $b_j(j=1,2\cdots)$ 同时成立(与∧)时,则该元组属于反例类 N。

1.3　知识表示

数据挖掘获取的规则知识一般表示形式

Condition(条件)→Result(结论)　(3)

2　可拓数据挖掘概念与理论

2.1　可拓数据挖掘概念

可拓学是利用可拓变换,即从变化的角度使假命题变为真命题,把不可知问题变为可知问题,把不可行问题转换为可行问题。可拓数据挖掘是在数据挖掘获得的静态知识的基础上,通过可拓变换,获取变化知识,即含可拓变换的规则知识,或可拓变换与相应的传导变换之间的规则知识。

2.2　可拓数据挖掘定理

定理 1 对于两类规则:

$\wedge a_i \rightarrow P$

$\wedge b_j \rightarrow N$

若存在条件的可拓变换 $T_{条件}$:

$$T_{条件}(\wedge b_j)=\wedge a_i \tag{4}$$

说明:对于 $j \geq i$,有 $T(b_k)=a_k, k=1,2,\cdots,i$

并存在结论的可拓变换 $T_{结论}$ (它为 $T_{条件}$ 的传导变换,即 $T_{条件} \rightarrow T_{结论}$):

$$T_{结论}(N)=P \tag{5}$$

则成立可拓变换规则知识(变化知识):

$$T(\wedge b_j)=\wedge a_i \quad T(N)=P \tag{6}$$

即:if $T(\wedge b_j)=\wedge a_i$ then $T(N)=P$ (7)

定理 2 对于两条同类规则:

$$A \rightarrow P \tag{8}$$

$$C \wedge B \rightarrow P \tag{9}$$

若存在可拓变换:

基金项目:国家博士点基金资助项目

作者简介:黄金才,男,博士在读博士后,研究方向:数据挖掘与智能决策技术。陈文伟,男,博导,教授,研究方向:数据挖掘,数据仓库与决策支持系统。

$T(B)=A$ (10)

则成立：可拓变换规则知识：

$T(B)=A\rightarrow P$ (11)

即 if $T(B)=A$ then P (12)

2.3 可拓数据挖掘获取的知识

可拓数据挖掘是在数据挖掘的基础上，通过可拓变换，得到可拓变换规则知识（变化知识），即从条件的可拓变换 $T_{条件}(T_{Condition})$ 和结论的可拓变换 $T_{结论}(T_{Result})$，获得可拓变换规则知识：

$T_{Condition}\rightarrow T_{Result}$ (13)

或 $T_{condition}\rightarrow T_{Result}$ (14)

3 可拓数据挖掘实例

3.1 “脑血栓”病与“脑出血”病发生变化的可拓数据挖掘实例

（1）通过数据挖掘获取规则知识

从“脑出血”和“脑血栓”两种疾病的大量实例数据库中，通过数据挖掘的遗传算法可以获取两种疾病独立诊断的规则知识。

A.脑出血和 B.脑血栓的病例判断，应当考虑如下几个方面的特征（属性）：

R_1 表示病人 N 的既往病史分物元，包括高血压病史 c_{11}（量值为 v_{11}=有 or 无）；动脉硬化病史 c_{12}（量值为 v_{12}=有 or 无）。记为：

$R_{11}=(N,c_{11},有)$，$\bar{R}_{11}=(N,c_{11},无)$；$R_{12}=(N,c_{12},有)$，$\bar{R}_{12}=(N,c_{12},无)$

以下物元取不同量值时的记法同上。

R_2 表示病人 N 的起病方式分物元（v_2=快 or 慢）；

R_3 表示病人 N 的局部症状分物元，包括：

c_{31} 偏瘫症状（v_{31}=是 or 否）；c_{32} 瞳孔不等大症状（v_{32}=是 or 否）；c_{33} 呕吐症状（v_{33}=是 or 否）；c_{34} 两便失禁情况（v_{34}=是 or 否）；c_{35} 语言障碍症状（v_{35}=是 or 否）；c_{36} 意识障碍症状（v_{361}=无，v_{362}=深度，v_{363}=轻度）；

R_4 表示病人 N 的病理反射情况分物元（v_4=阳 or 阴）；

R_5 表示病人 N 的膝腱反射情况分物元（v_{51}=无，v_{52}=活跃，v_{53}=不活跃），记：

$R_{51}=(N,c_5,无)$，$R_{52}=(N,c_5,活跃)$，$R_{53}=(N,c_5,不活跃)$

R_6 表示病人 N 的病情发展速度分物元（v_6=快 or 慢）；

案例的物元表示为：

$$R=\begin{bmatrix} N,病史\ c_1, & v_1\\ 起病方式\ c_2, & v_2\\ 局部症状\ c_3, & v_3\\ 病理反射\ c_4, & v_4\\ 膝腱反射\ c_5, & v_5\\ 病情发展速度\ c_6, & v_6 \end{bmatrix}=\begin{bmatrix} R_1\\ R_2\\ R_3\\ R_4\\ R_5\\ R_6 \end{bmatrix}$$

$$R_1=\begin{bmatrix} N,高血压病史\ c_{11}, & v_{11}\\ 动脉硬化病史\ c_{12}, & v_{12} \end{bmatrix}=\begin{bmatrix} R_{11}\\ R_{12} \end{bmatrix}$$

$$R_3=\begin{bmatrix} N,偏瘫症状\ c_{31}, & v_{31}\\ 两瞳孔症状\ c_{32}, & v_{32}\\ 呕吐症状\ c_{33}, & v_{33}\\ 两便症状\ c_{34}, & v_{34}\\ 语言症状\ c_{35}, & v_{35}\\ 意识症状\ c_{36}, & v_{36} \end{bmatrix}=\begin{bmatrix} R_{31}\\ R_{32}\\ R_{33}\\ R_{34}\\ R_{35}\\ R_{36} \end{bmatrix}$$

上面是从六个方面 12 个特征来识别诊断患者到底得的是“脑出血”还是“脑血栓”。设：

$R_A=(N,病症,脑出血)$，$R_B=(N,病症,脑血栓)$

我们从大量的“脑出血”和“脑血栓”病人的病例中，进行数据挖掘的遗传算法学习后，获得的主要 7 条规则（具体数据挖掘过程从略）见表 1。

表 1 获得的主要 7 条规则

1	$R_{11}\wedge R_{32}\wedge R_{53}\rightarrow R_A$
2	$R_{32}\wedge R_{35}\rightarrow R_A$
3	$R_{11}\wedge R_2\wedge R_{362}\rightarrow R_A$
4	$R_{11}\wedge R_6\rightarrow R_A$
5	$R_{11}\wedge R_{12}\wedge \bar{R}_2\rightarrow R_B$
6	$R_{12}\wedge \bar{R}_6\rightarrow R_B$
7	$R_{12}\wedge R_{361}\rightarrow R_B$

（2）通过可拓数据挖掘获取可拓变换规则知识（变化知识）

不少脑血栓的患者会转换成脑出血。很多医生会按原有的观念，在病人的病情发生转变后，仍按原方案治疗，诊断失误，造成病情加重。这是因为：①脑血栓病人的治疗是要疏通血管；②脑出血病人的治疗是要堵塞血管。这是两种截然相反的治疗方法。当病人已由脑血栓转换成脑出血后，仍用脑血栓的治疗方式治脑出血，即通血管治疗，结果使已脑出血的病人更大范围的出血，造成大出血，甚至死亡。

问题是：如何诊断脑血栓的病人出现了脑出血？

根据可拓数据挖掘定理 1 可知：将规则 7 的前提条件作可拓变换成规则 5 的前提条件，即前提的可拓变换：

$T_1(R_{11}\wedge R_{12}\wedge \bar{R}_2\wedge R_{361})=R_{11}\wedge R_2\wedge R_{362}$

按可拓数据挖掘定理 1，它与结论的可拓变换构成可拓变换知识 1（变化知识）为：

T（有动脉硬化∧起病方式慢∧无意识障碍）=起病方式快∧有深度意识障碍：

$\rightarrow T(R_B)=R_A$

该变化规则知识表明，当发现病人由起病方式慢变成起病方式快，同时无意识障碍变成有深度意识障碍，就应该诊断该病人已经由“脑血栓”变成了“脑出血”。治疗方式就应由“脑血栓”的治疗方法变成治疗“脑出血”的方法，这是两种相反的治疗方法。若仍然用“脑血栓”的治疗方法治疗“脑出血”，将会快速加重“脑出血”症状。这条变化知识对医生来讲是极其重要的。

此例说明变化知识的价值和重要性。还可以得到其它可拓变换知识，在此从略。

3.2 气候发生变化影响打高尔夫球的判断实例

气候包括四个属性和属性值：

天气：晴、多云、雨

气温：冷、适中、热

湿度：高、正常

（下转 25 页）

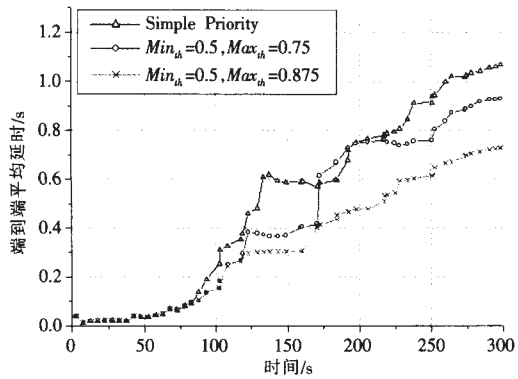

图 7　重负载时端到端平均延时比较

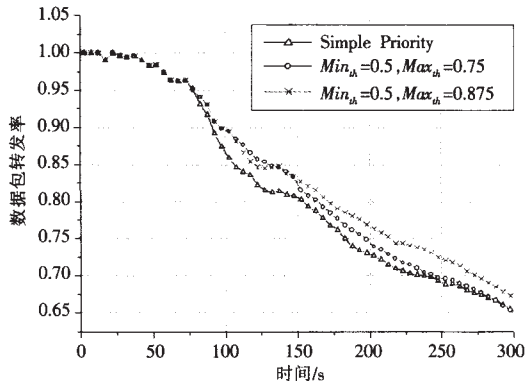

图 8　重负载时数据包成功转发率比较

个移动节点的随机拓扑，仿真的网络范围为 1 000m×750m，节点的最大移动速度为 10m/s，两次连续移动间的平均停顿时间为 2s，路由协议采用 AODV，业务流为随机产生的 2kbps 的 CBR 数据流，网络中存在的最大连接数为分别为 15、20、25，仿真持续时间为 300s，缓冲区大小为 50。

在性能评价中，我们采用所有连接的端到端平均延时以及数据包的成功转发率作为评价指标。由图中可以看出，在不同的 Min_{th} 和 Max_{th} 组合下 DPQS 机制较简单优先权算法性能均有所提高，而在 Min_{th} 取 0.5，Max_{th} 取 0.875 时效果更为明显。在中负载和重负载情况下，随着网络中通信量的增加其可使网络延时最多降低约 30%，数据包的成功转发率最大也可提高近 5%。

6　结束语

MANET 环境下的队列调度机制的关键是解决如何在保证快速建立路由的前提下，降低整个网络中数据包的传输延时。另外，由于节点频繁移动引起的路由切换会造成网络中路由数据包的突发现象，也会导致网络传输延时的增加。本文的 DPQS 机制，提出在网络层根据节点的负载情况动态调整路由消息和普通数据包的转发优先权。仿真结果表明 DPQS 机制可以有效地降低网络传输延时，并在一定程度上可以提高网络的吞吐量，可以有效地配合 MANET 路由协议提高网络性能。

（收稿日期：2006 年 3 月）

参考文献

1.C Perkins, E Royer, S Das.Ad-hoc on demand distance vector(aodv) routing[S].IETF Internet Draft, 2003

2.Johnson D, Maltz D, Hu Y et al.The dynamic source routing protocol (DSR) for mobile Ad Hoc networks.http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-09.txt, 2003

3.K Chandran, S Raghunathan, S Venkatesan et al.A feedback-based scheme for improving TCP performance in ad hoc wireless networks[J].IEEE Personal Communications Magazine, 2001; 8(1): 34~39

4.J Liu, S Singh.ATCP: TCP for mobile ad hoc networks1[J].IEEE Journal on Selected Areas in Communications, 2001; 19(7): 1300~1315

5.冯彦君，孙利民，钱华林等.MANET 中 TCP 改进研究综述[J].软件学报，2005; 16(3): 434~444

6.K Tang, M Gerla.Fair sharing of MAC under TCP in wireless ad hoc networks[C].In: Proceedings of IEEE MMT'99, 1999-10

7.S Xu, T Saadawi.Does the IEEE 802.11 MAC protocol work well in multi-hop wireless ad hoc networks?[J].IEEE Communications Magazine, 2001; 39(6)

8.Xiao Long Huang, Brahim Bensaou.On Max-min Fairness and Scheduling in Wireless Ad-Hoc Networks: Analytical Framework and Implementation[C].In: IEEE/ACM MobiHOC, Long Beach, CA, USA, 2001

9.ns-2.http://www.isi.edu/nsnam/ns/

10.陆传赉.排队论[M].北京：北京邮电学院出版社，1993

（上接 8 页）

风：有风、无风

通过大量各种气候条件下打高尔夫球的实例，可以打（P 类）和不可以打（N 类），利用 ID3 决策树的数据挖掘方法，获取以下决策树知识，见图 1。

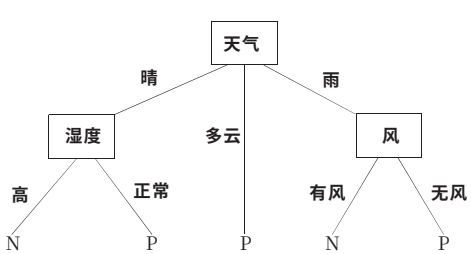

图 1　决策树知识

利用可拓数据挖掘定理 2，可以获得可拓变换规则知识如下：

(1) if T(天气=雨)=(天气=多云) then 类别=P

(2) if T(天气=雨)=(天气=晴)∧(湿度=正常) then 类别=P

(3) if T(天气=雨)=(天气=晴)∧(湿度=高) then 类别=N

以上变化规则知识说明：在天气发生变化后，需要根据变化的知识来决定是否能打高尔夫球。此例也可如上例用物元进行形式化表示，此略。（收稿日期：2006 年 1 月）

参考文献

1.陈文伟等.数据挖掘技术[M].北京：北京工业大学出版社，2002-12

2.蔡文，杨春燕，何斌.可拓逻辑初步[M].北京：科学出版社，2003-11

3.Agralwal R, Imielinski T.Mining Association Rules Between Sets of Items in Large Database[C].In: Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data, Washington, 1993: 207~216

4.Agralwal R, Srikant R.Fast Algorithm For Mining Association Rules[C].In: Proceedings of the 1994 International Conference on Very Large Data Base, Santiago, Chile, 1994: 487~499