第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)04—0053—03 文献标识码：A 中图分类号：TP311.13

基于可分辨关系的知识约简

陈鑫影，邱占芝
(大连交通大学软件学院，大连 116028)

摘 要：为寻求求解约简的有效方法，从而有效处理大规模数据，并减少后续挖掘算法在时间和空间上的压力，基于粗糙集理论提出可分辨关系的概念，并在此基础上定义对象差异矩阵、分辨约简集和分辨核心集等概念，证明划分约简这一传统知识约简与分辨约简的一致性，讨论其他概念间的关系，并给出相关的定理和等价命题。通过理论论证和示例分析，可以获知基于可分辨关系的属性约简的有效性和可行性。

关键词：数据挖掘；粗糙集；知识约简；可分辨关系

Knowledge Reduction Based on Distinguishable Relation

CHEN Xin-ying, QIU Zhan-zhi
(Software Technology Institute, Dalian Jiaotong University, Dalian 116028)

【Abstract】 It is meaningful to approach new ways for achieving knowledge reduction in information systems with the huge volume of data. This paper gives distinguishable relation based on rough set theory. Against the new concepts, distinguishable matrix, distinguishable reduction and distinguishable core are presented. The main objective of this paper is to find and prove the relationship between the classical types of knowledge reduction, such as partition reduction and distinguishable reduction. The result shows, that distinguishable reduction is equivalent to partition reduction under all conditions, is easy to conclude. The relationships among other conceptions proposed are discussed. The judgment theorems and equivalent definitions with respect to these new concepts are obtained. New concepts and ways are introduced to figure out knowledge reduction in information systems, and the knowledge reductions based distinguishable relation is meaningful both in the theory and in applications.

【Key words】 data mining; rough set; knowledge reduction; distinguishable relation

1 概述

粗糙集理论(rough sets)由波兰科学家 Pawlak 在 1982 年首先提出[1-2]，是一种新的处理含糊性(vagueness)和不确定性(uncertainty)问题的数学工具，它从新的视角对知识进行了定义，把知识看作是关于论域的划分，认为知识是有粒度的[3]。粗糙集理论主要用于知识约简及知识依赖性的分析[4-5]，知识约简可以通过删除冗余知识来得到知识库的本质信息，从而提取深层次的用于描述知识库整体特征以及发展趋势的预测内容，并进一步完成分类以及其他工作，可有效应对大规模数据。

本文提出了可分辨关系的概念，论证了划分协调集与一致分辨集、划分约简集与分辨约简集的一致性，分析了对象差异矩阵的性质，并给出了基于可分辨关系的属性约简的具体操作方法。

2 信息系统与可分辨关系

定义 1(信息系统) 四元组 $S=(U, A, V, F)$是一个知识表达系统(信息系统)，其中，$U=\{x_1, x_2,\cdots, x_n\}$为对象的非空有限集合，称为论域；$A=\{a_1, a_2,\cdots, a_m\}$为属性的非空有限集合，称为属性集；$V=\cup V_k$，对属性集中的任一属性 $a_k\in A(k\leq m)$，V_k 看作其值域；$F=\{f_k: U\to V_k(k\leq m)\}$，即对给定对象 $x_i(i\leq n)$，$f_k(x_i)$赋予对象 x_i 在属性 a_k 下的属性值。知识表达系统也称为信息系统。通常也用 $S=(U, A)$来代替 $S=(U, A, V, F)$。

定义 2(可分辨关系) 设信息系统 $S=(U, A)$，设 $P\subseteq A$ 且 $P\neq\varnothing$，定义由属性子集 P 导出的二元关系：

$$\mathrm{DIS}_P=\{(x_i,x_j)\,|\,(x_i,x_j)\in U\times U,\ \exists a_k\in P\ f_k(x_i)\neq f_k(x_j)\}$$

称 DIS_P 为由属性集 P 推导出的可分辨关系。

类似的，定义由属性 $a_k\in A$ 导出的可分辨关系为

$$\mathrm{DIS}_{\{a_k\}}=\{(x_i,x_j)\,|\,(x_i,x_j)\in U\times U,\ f_k(x_i)\neq f_k(x_j)\}$$

简记为 DIS_{a_k}。

定义 3(可分辨对象单元集) 若$(x_i, x_j)\in\mathrm{DIS}_P$，则称 x_i和 x_j 是 P 可分辨的，即依据属性集 P 可将 x_i和 x_j 区分开，并将(x_i, x_j)称为可分辨对象单元。将由属性集 P 推导出的可分辨关系对应的集合称为属性集 P 的可分辨对象单元集，用 $\mathrm{DIS}_P(U)$表示。类似的，将由属性 $a_k\in A$ 推导出的可分辨关系对应的集合称为属性 a_k的可分辨对象单元集，用 $\mathrm{DIS}_{a_k}(U)$ 表示。

定义 4(对象差异矩阵) 设信息系统 $S=(U, A)$，其中论域 $U=\{x_1, x_2, \cdots, x_n\}$，属性集合$A=\{a_1, a_2,\cdots, a_m\}$，即$|U|=n$, $|A|=m$，将 S 的对象差异矩阵 $\bar{\boldsymbol{M}}$ 定义为一个具有 n^2 行、m 列的矩阵，其 k 列(k=1, 2,…,m)的元素定义为

$$\bar{m}_k=\{(x_i,x_j)\,|\,(x_i,x_j)\in U\times U,\ a_k\in A, f_k(x_i)\neq f_k(x_j)\}$$

即 $\bar{m}_k$ 为属性 a_k 的可分辨对象单元集。

基金项目：辽宁省自然科学基金资助项目(20072157)；辽宁省教育厅高校科研计划基金资助项目(20060107, 2009A132)
作者简介：陈鑫影(1979－)，女，讲师、硕士，主研方向：数据挖掘，粗糙集理论；邱占芝，教授、博士
收稿日期：2009-09-29 **E-mail：**chenxy1979@163.com

定理 1 设信息系统 $S=(U, A)$，若存在 $P\subseteq A$ 且 $P\neq\varnothing$，则 DIS_P 具有如下性质：

(1)若 $P_1\subseteq P_2\subseteq A$，则 $DIS_{P_1}\subseteq DIS_{P_2}\subseteq DIS_A$ ；

(2) $DIS_P=\bigcup_{a_k\in P} DIS_{a_k}$。

证明:

(1)对于$\forall(x_i, x_j)\in DIS_{P_1}$，由定义 2 可知必然$\exists a_k\in P_1$，使得 $f_k(x_i)\neq f_k(x_j)$，由于 $P_1\subseteq P_2$，故 $a_k\in P_2$，于是由定义 2 可知$(x_i, x_j)\in DIS_{P_2}$ 成立，因此 $DIS_{P_1}\subseteq DIS_{P_2}$。同理，易见 $DIS_{P_2}\subseteq DIS_A$。因此若 $P_1\subseteq P_2\subseteq A$，则 $DIS_{P_1}\subseteq DIS_{P_2}\subseteq DIS_A$。

(2)由定义 2 可知对于$\forall(x_i, x_j)\in\bigcup_{a_k\in P} DIS_{a_k}$，必然$\exists a_k\in P$，使得$(x_i, x_j)\in DIS_{a_k}$，即$\exists a_k\in P$，使得 $f_k(x_i)\neq f_k(x_j)$，于是由定义 2 可知 $(x_i, x_j)\in DIS_P$，则 $\bigcup_{a_k\in P} DIS_{a_k}\subseteq DIS_P$。同理，易见 $DIS_P\subseteq\bigcup_{a_k\in P} DIS_{a_k}$。因此 $DIS_P=\bigcup_{a_k\in P} DIS_{a_k}$。

定理 2 设信息系统 $S=(U, A)$，对象差异矩阵为 $\bar{\boldsymbol{M}}$，对于 $\forall x_i, x_j, x_k\in U$，$\forall a_l\in A$，对象差异矩阵 $\bar{\boldsymbol{M}}$ 具有如下性质：

(1)$(x_i, x_i)\notin\bar{m}_l$。

(2)若$(x_i, x_j)\notin\bar{m}_l$，则$(x_j, x_i)\notin\bar{m}_l$；若$(x_i, x_j)\in\bar{m}_l$，则$(x_j, x_i)\in\bar{m}_l$。

(3)若$(x_i, x_j)\notin\bar{m}_l$，$(x_i, x_k)\in\bar{m}_l$，则$(x_j, x_k)\in\bar{m}_l$。

证明：

(1)与(2)易证。

(3)$(x_i, x_j)\notin\bar{m}_l$ 等价于 $f_l(x_i)=f_l(x_j)$，$(x_i, x_k)\in\bar{m}_l$ 等价于 $f_l(x_i)\neq f_l(x_k)$，由此可知 $f_l(x_j)\neq f_l(x_k)$，因此$(x_j, x_k)\in\bar{m}_l$。

3 划分约简与分辨约简

定义 5(划分约简集) 设$S=(U, A)$为信息系统,若存在 $P\subseteq A$ 且 $P\neq\varnothing$，使 $R_P=R_A$，则称 P 为划分协调集[2]。若 P 为划分协调集，且 P 的任何真子集均不是划分协调集，则称 P 为信息系统 S 的划分约简集[2]。

定义 6(分辨约简集) 设信息系统 $S=(U, A)$，若存在 $P\subseteq A$ 且 $P\neq\varnothing$，使 $DIS_P=DIS_A$，则称 P 为一致分辨集。若 P 为一致分辨集，且对于$\forall Q\subset P$，满足 $DIS_Q\neq DIS_A$，则称 P 为信息系统 S 的分辨约简集。

定义 7(分辨核心集) 设信息系统 $S=(U, A)$,设信息系统 S 存在 r 个分辨约简集，$P_k(k=1, 2, \cdots, r)$，定义信息系统的分辨核心集为

$$\overline{CORE}=\bigcap_{1\le k\le r} P_k$$

定理 3 设信息系统 $S=(U, A)$，若存在 $P\subseteq A$ 且 $P\neq\varnothing$，则以下命题等价：

(1)P 是一致分辨集($DIS_P=DIS_A$)；

(2)$DIS_A\subseteq DIS_P$。

证明：

(2)⇒(1)

由定理 1 中性质(1)易知若 $P\subseteq A$，则 $DIS_P\subseteq DIS_A$，由于 $DIS_A\subseteq DIS_P$，因此 $DIS_P=DIS_A$。

(1)⇒(2)

由 $DIS_P=DIS_A$，易知 $DIS_P\subseteq DIS_A$，$DIS_A\subseteq DIS_P$。

定理 4 设信息系统 $S=(U, A)$，则分辨约简集必然存在。

证明：

(1)若对于$\forall a_k\in A$，满足 $DIS_{A-\{a_k\}}\neq DIS_A$，即对于$\forall P\subset A$，$DIS_P\neq DIS_A$，则 A 为分辨约简集。

(2)若$\exists a_k\in A$，使得 $DIS_{A-\{a_k\}}=DIS_A$，则设 $P=A-\{a_k\}$，若$\forall a_l\in P$，满足 $DIS_{P-\{a_l\}}\neq DIS_A$，则由定义 6 易知 P 为分辨约简集。

若$\exists a_l\in P$，满足 $DIS_{P-\{a_l\}}=DIS_A$，则设 $P'=A-\{a_k\}-\{a_l\}$，重复(2)。由于 A 为属性的非空有限集合，即$|A|$为有限，因此通过有限步必然$\exists P''\subset A$,使得 $DIS_{P''}=DIS_A$，且对于$\forall a\in P''$，满足 $DIS_{P''-\{a\}}\neq DIS_A$，此时 P'' 为分辨约简集。可见对于信息系统 $S=(U, A)$，分辨约简集必然存在。

定理 5 设信息系统 $S=(U, A)$，则系统 S 的划分协调集和一致分辨集是一致的，划分约简集和分辨约简集是一致的。

证明：

设 $P\subseteq A$ 且 $P\neq\varnothing$，P 为划分协调集等价于 $R_P=R_A$；P 为一致分辨集等价于 $DIS_P=DIS_A$。

P 为划分约简集等价于 $R_P=R_A$，且对于$\forall Q\subset P$，满足 $R_Q\neq R_A$，即 $R_P\subseteq R_A$成立，但 $R_Q\subseteq R_A$不成立。

P 为分辨约简集等价于 $DIS_P=DIS_A$，且对于$\forall Q\subset P$，满足 $DIS_Q\neq DIS_A$，即 $DIS_A\subseteq DIS_P$成立，但 $DIS_A\subseteq DIS_Q$不成立。

(1) 对于 $\forall(x_i, x_j)\in U\times U$，$DIS_A\subseteq DIS_P$ 等价于若 $(x_i, x_j)\in DIS_A$，则 $(x_i, x_j)\in DIS_P$，即若 $[x_i]_A\neq[x_j]_A$，必然有 $[x_i]_P\neq[x_j]_P$。

反之若$(x_i, x_j)\notin DIS_P$，必然有$(x_i, x_j)\notin DIS_A$，即对于$\forall a_k\in P$，当 $f_k(x_i)=f_k(x_j)$时，必然有$\forall a_l\in A$，$f_l(x_i)=f_l(x_j)$，也即当 $[x_i]_P=[x_j]_P$ 时，必然有$[x_i]_A=[x_j]_A$，从而当 $x\in U$, $[x_i]_P=[x_j]_P=[x]_P$, $x_i, x_j\in[x]_P$ 时，必然有 $[x_i]_A=[x_j]_A=[x]_A$, $x_i, x_j\in[x]_A$。则证 $[x]_P\subseteq[x]_A$, $R_P\subseteq R_A$，由于 $R_A\subseteq R_P$ 总成立，因此 $R_P=R_A$。

因此，$DIS_P=DIS_A$等价于 $R_P=R_A$，即划分协调集和一致分辨集是一致的。

(2) 由 (1) 易知，对于 $\forall Q\subset P$，若 $DIS_P=DIS_A$，且 $DIS_Q\neq DIS_A$，必然有 $R_P=R_A$，且 $R_Q\neq R_A$，故若 P 为分辨约简集，则必为划分约简集。同理易证若 P 为划分约简集，则必为分辨约简集。

因此，划分约简集和分辨约简集是一致的。

定理 6 设信息系统 $S=(U, A)$，对象差异矩阵为 $\bar{\boldsymbol{M}}$，$P\subseteq A$ 且 $P\neq\varnothing$，P 为一致分辨集当且仅当对于$\forall(x_i, x_j)\in U\times U$，若$\exists a_k\in A$，使得$(x_i, x_j)\in\bar{m}_k$，则必然$\exists a_l\in P$，使得$(x_i, x_j)\in\bar{m}_l$。

证明：

P 为一致分辨集等价于 $DIS_A\subseteq DIS_P$，等价于$\forall(x_i, x_j)\in DIS_A$时，$(x_i, x_j)\in DIS_P$，也即若$\exists a_k\in A$，使得 $f_k(x_i)\neq f_k(x_j)$，必然$\exists a_l\in P$，使得 $f_l(x_i)\neq f_l(x_j)$，于是由定义 4 可知，若$(x_i, x_j)\in\bar{m}_k$，则必然$(x_i, x_j)\in\bar{m}_l$，故得证。

定理 7 设信息系统 $S=(U, A)$, $(x_i, x_j)\in U\times U$, $a_k\in A$，则以下命题等价：

(1)$a_k\in\overline{CORE}$；

(2)$\exists(x_i, x_j)\in\bar{m}_k$，对于$\forall a_l\in A-\{a_k\}$, $(x_i, x_j)\notin\bar{m}_l$；

(3) $DIS_{A-\{a_k\}}\neq DIS_A$。

证明：

(1)⇒(2)

若(2)不成立，即对于$\forall(x_i, x_j)\in\bar{m}_k$ 时，必然$\exists a_l\in A-\{a_k\}$，使得$(x_i, x_j)\in\bar{m}_l$，等价于当$(x_i, x_j)\in DIS_{a_k}$ 时，必然$\exists a_l\in A-\{a_k\}$，使得$(x_i, x_j)\in DIS_{a_l}$，即当$(x_i, x_j)\in DIS_A$ 时，$(x_i, x_j)\in DIS_{A-\{a_k\}}$ 成

立，因此 $\mathrm{DIS}_A \subseteq \mathrm{DIS}_{A-\{a_k\}}$，于是 $\mathrm{DIS}_A = \mathrm{DIS}_{A-\{a_k\}}$，从而必然 $\exists P \subseteq A-\{a_k\}$使 P 为分辨约简集，由于 $a_k \notin P$，由定义 7 易知 $a_k \notin \overline{CORE}$，这与 $a_k \in \overline{CORE}$ 相矛盾，故(2)成立。

(2)⇒(3)

当$(x_i, x_j) \in \bar{m}_k$ 时，对于$\forall a_l \in A-\{a_k\}$，$(x_i, x_j) \notin \bar{m}_l$ 成立，等价于当 $f_k(x_i) \neq f_k(x_j)$时，对于$\forall a_l \in A-\{a_k\}$，必然有 $f_l(x_i)=f_l(x_j)$，即当$(x_i, x_j) \in \mathrm{DIS}_{a_k}$ 时，则必然$(x_i, x_j) \notin \mathrm{DIS}_{a_l}$，从而当$(x_i, x_j) \in \mathrm{DIS}_A$时，$(x_i, x_j) \notin \mathrm{DIS}_{A-\{a_k\}}$。虽然 $A-\{a_k\} \subseteq A$，$\mathrm{DIS}_{A-\{a_k\}} \subseteq \mathrm{DIS}_A$成立，但 $\mathrm{DIS}_A \subseteq \mathrm{DIS}_{A-\{a_k\}}$ 不成立，因此 $\mathrm{DIS}_{A-\{a_k\}} \neq \mathrm{DIS}_A$。

(3)⇒(1)

若(1)不成立，即 $a_k \notin \overline{CORE}$，由定义 7 易知 $a_k \notin \bigcap_{1 \le k \le r} P_k$，即$\exists P_k$，使得 $a_k \notin P_k$，也即必然$\exists P_k \subseteq A-\{a_k\}$使 P_k为分辨约简集，于是 $\mathrm{DIS}_A = \mathrm{DIS}_{A-\{a_k\}} = \mathrm{DIS}_{P_k}$，与(3)矛盾，故(1)成立。

例　设有信息系统 $S=(U, A)$,其中论域 $U=\{x_1, x_2, x_3, x_4\}$，条件属性集 $A=\{a_1, a_2, a_3\}$，利用数据表格表示，见表 1。

表 1　信息系统

U	a_1	a_2	a_3
x_1	1	1	2
x_2	1	2	1
x_3	1	2	1
x_4	2	2	1

首先，由定义 2 得到 DIS_{a_1}, DIS_{a_2} 和 DIS_{a_3}，则由定理 1 性质(2)可以得到 DIS_A。而后由定义 4 得到 S 的对象差异矩阵，其中，若$(x_i, x_j) \in \mathrm{DIS}_{a_k}$，则矩阵$(i-1)\cdot n+j$ 行($|U|=n=4$)k 列处以“1”表示；反之以“0”表示。此后由定理 2 得到简化的对象差异矩阵，见表 2。此时取 $P_1=\{a_1, a_2\}$，由定理 1 性质(2)和定义 6 可知 P_1 为 S 的分辨约简集。取 $P_2=\{a_1, a_3\}$，由定理 3 和定义 6 可知 P_2 为分辨约简集。取 $P_3=\{a_2, a_3\}$，由定理 6 可知 P_3 不是分辨约简集。则 S 的分辨约简集为 P_1,P_2(或者亦可先求取划分约简集，然后由定理 5 得到分辨约简集)。此后由定义 7 或者定理 7 均能求取 S 的分辨核心集 $\overline{CORE} = \{a_1\}$。

表 2　简化的对象差异矩阵

U	a_1	a_2	a_3
(x_1, x_2)	0	1	1
(x_1, x_3)	0	1	1
(x_1, x_4)	1	1	1
(x_2, x_3)	0	0	0
(x_2, x_4)	1	0	0
(x_3, x_4)	1	0	0

4　结束语

本文提出可分辨关系的概念，并针对信息系统提出了相关概念，同时论证了划分约简与分辨约简的一致性。基于示例可以看出与划分约简相比，基于可分辨关系的属性约简方法可将相关的逻辑运算转换成矩阵运算，故可以达到简化计算、降低时空开销、有效处理海量数据的目的。本文为后续研究具体的属性约简算法打下了理论基础，为属性约简提供了新的解决途径，未来将进一步研究多关系表约简和动态属性约简问题。

参考文献

[1] Pawlak Z. Rough Sets[J]. Internetional Joural of Computer and Information Sciences, 1982, 11(5): 341-356.

[2] Pawlak Z. Rough Sets——Theoretical Aspects of Reasoning About Data[M]. Dordreche, Holland: Kluwer Academic Publishers, 1991.

[3] 何　明. 一种基于粒度的粗糙聚类分析方法[J]. 计算机工程, 2008, 34(8): 203-204.

[4] Pawlak Z. Rough Set Theory and Its Application to Data Analysis[J]. Cybernetics and Systems, 1998, 45(9): 661-668.

[5] 瞿彬彬, 卢炎生. 基于粗糙集的快速属性约简算法研究[J]. 计算机工程, 2007, 33(11): 7-9.

编辑　任吉慧

(上接第 52 页)

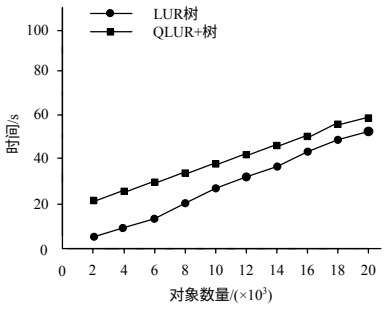

图 9　更新性能对比

5　结束语

本文提出 QLUR+树索引结构，主要有以下几方面的改进：(1)QLUR+树采用基于 R+树结构，最大限度地减少无效查询，提高索引效率；(2)采用懒惰更新和扩充与收缩 MBR 方法，减少对象删除和重插入，更新性能得到了保证；(3)四叉树存储移动对象过去一段时间的位置信息；(4)引入了附加索引结构 F-L，可直接访问移动对象当前及最近历史信息。经实验及分析得出此结构具有很高的查询性能，整体优势明显，且支持大型数据库索引。QLUR+树基于四叉树和 R+树结构，实现较为简单，在需要频繁查询检索领域具有重要的应用价值。空间数据库的空间查询处理是数据库的关键技术之一，利用空间索引进行高效的空间查询处理研究将具有重要的理论与应用价值[4]。下一步的工作将研究空间查询处理中的空间连接和最近邻查询问题。

参考文献

[1] 王国仁, 黄健美. 基于最大间隙空间映射的高维数据索引技术[J]. 软件学报, 2007, 18(6): 1419-1428.

[2] 王平根, 周脚根. 一种混合的时空数据库索引机制[J]. 计算机科学, 2007, 34(9): 103-106.

[3] Theodoridis Y. On the Generation of Spatiotemporal Datasets[C]// Proc. of the 6th Int'l Symp. on Spatial Databases. [S. l.]: IEEE Press, 1999.

[4] 张明波, 陆　峰. R 树家族的演变和发展[J]. 计算机学报, 2005, 28(3): 229-230.

编辑　陈　文