

一种专家系统知识获取时的属性值约简算法

彭霞,朱萍,任永昌

(渤海大学信息科学与技术学院,辽宁锦州 121013)

摘要:知识获取是构造专家系统的“瓶颈”,提供准确的推理知识是进行科学决策的关键。文中运用粗糙集理论,研究对决策表中每条记录的冗余条件属性值进行筛选并删除的属性值约简算法。首先研究属性值约简的理论基础,包括知识表示和知识约简与核两个方面;其次研究知识获取方式与知识获取过程;然后研究属性值约简算法,通过两个定义描述约简算法的基础上,给出了约简算法的5个步骤;最后以城市物流中心选址为例,运用属性值约简算法及其步骤,对决策表属性值进行约简。结果表明,属性值约简实现了决策表的最简化,突出了关键属性及其关键属性值对决策的影响。

关键词:专家系统;知识获取;属性值约简算法;粗糙集理论

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2013)04-0154-05

doi:10.3969/j.issn.1673-629X.2013.04.038

An Attribute Value Reduction Algorithm of Expert System Knowledge Acquisition

PENG Xia, ZHU Ping, REN Yong-chang

(College of Information Science and Technology, Bohai University, Jinzhou 121013, China)

Abstract: To construct an expert system, knowledge acquisition is the "bottleneck" problem, and the key to scientific decision-making is accurate inference knowledge. In this paper, it will use rough sets theory to attribute value reduction algorithms that filter and remove the redundant condition attribute value of decision-making table's each record. First, research the theoretical basis of attribute value reduction, including knowledge representation and knowledge reduction and core both. Then, research knowledge acquisition ways and knowledge acquisition processes. Third, research the attribute value reduction algorithms, through the description of two definitions the reduction algorithms is given the 5 steps. Finally, take the city logistics center location as example, use attribute value reduction algorithms and its steps for the value of decision table attribute reduction. The results show that attribute value reduction achieves the most streamlined decision-making table, highlighting the key attributes and key decision-making of property value.

Key words: expert system; knowledge acquisition; attribute value reduction algorithms; rough sets theory

0 引言

随着信息技术的不断发展和普及,人类积累的数据量正在以指数速度增长。但与这些“海量”数据相比,人们分析数据的能力以及从中获取知识的能力都存在着相当大的差距,因此形成了“数据过剩”而又“信息匮乏”的被动局面。如何科学、合理、有效、正确地利用这些数据,以产生对人们有价值的信息和知识,已成为学术界面临的必须解决的挑战性问题^[1]。粗糙集理论从集合的视角对知识进行定义,把知识看作是论域的划分,构成一个信息系统,从而对知识进行分析和处理^[2]。

知识约简分为属性约简和属性值约简两个内容。决策表属性集中的每个属性并不都是必要的,通过属性约简,将决策表中对决策分类不必要的属性省略,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^[3]。但是,属性约简只是在一定程度上去掉了冗余信息,并不充分,需要进一步进行属性值约简。属性值约简是对某一条决策规则去掉某些值不改变整个系统的决策能力,本质就是决策规则去掉属性值后不会与其它规则产生冲突,即属性值完全相同而决策属性值却不同。属性值约简是属性约简的进一步深化,从而真正实现了决策表的最简化,同时更加突出了关键属性及其关键属性值对决策的影响^[4]。

1 理论基础

1.1 知识表示

知识是人们在长期的生活及社会实践中、科学研

收稿日期:2012-07-17;修回日期:2012-10-23

基金项目:2011 辽宁省科学事业公益基金(编号略)

作者简介:彭霞(1977-),女,讲师,硕士,从事计算机信息管理、软件项目开发技术研究。

究及实验中积累起来的,对客观世界的认识与经验,人们把实践中获得的信息关联在一起,就获得了知识。知识反映了客观世界中事物之间的联系,不同事物之间或者相事物间的不同关系形成了不同的知识。给知识下一个明确的定义很困难,不同的人有不同的理解。下面仅给出几个著名专家的看法:Feigenbaumy 认为,知识是经过消减、塑造、解释和转换的信息;Bernstein 认为,知识是由特定领域的描述、关系和过程组成的;Hayes-roth 认为,知识是事实、信念和启发式规则。对知识进行表示的过程就是把知识编码成为某种数据结构的过程^[5]。

知识表示是建立专家系统的基础,离开知识的具体表示,获取的知识就无法记录下来,也就无法建立知识库,进而建立在知识库上的推理也就无法进行。一般来说,同一知识有多种表示方式。表示方式的选择直接影响推理的效率以及知识的获取,因此,结合要解决的领域问题,将它与推理及知识获取方式结合起来考虑,选取最佳的知识表达方式非常重要。知识表达系统通过一定的方法从大量数据中发现有用的知识或决策规则^[6]。

知识表示系统可以表述为:

$$S = (U, A, V, f) \quad (1)$$

其中: U 表示对象的非空有限集合,称为论域;

A 表示属性的非空有限集合;

$V = \bigcup_{a \in A} V_a$, V_a 是属性 a 的值域;

$f: U \times A \rightarrow V$ 是一个信息函数,为每个对象的每个属性赋予一个信息值,即:

$$\forall a \in A, x \in U, f(x, a) \in V_a \quad (2)$$

式(1)也可以简化表示为:

$$S = (U, A) \quad (3)$$

若 A 可进一步划分为条件属性 C 和决策属性 D , 即 $A = C \cup D$, 则信息系统也可表示为:

$$S = (U, C, D, V, f) \quad (4)$$

在知识表达系统中,知识存储在知识库中,知识库中的数据在关系数据库中以二维表的形式存在,关系是笛卡尔积的有限子集,关系表中的每行对应一个元组,每列对应一个属性。在信息系统中,每个表存储一类对象(实体型),每行存储一个对象,每列存储对象的一个属性。

对于每个属性子集 $P \subseteq A$, 定义属性子集的不可分辨二元关系 $\text{ind}(P)$ 为:

$$\text{ind}(P) = \{ (x, y) \in U \times U, \forall a \in P, f(x, a) = f(y, a) \} \quad (5)$$

$\text{ind}(P)$ 是一个等价关系,满足:

$$\text{ind}(P) = \bigcap_{a \in P} \text{ind}(a) \quad (6)$$

任意知识库 $K = (U, R)$ 可用下面的方法指定一个

信息系统 $S = (U, A, V, f)$, 即当 $r \in R$, 且 $\frac{U}{R} = (X_1, X_2, \dots, X_k)$, 指定属性集 A 中的属性:

$$a_R: U \rightarrow V_{a_R}, V_{a_R} = \{1, 2, \dots, k\} \quad (7)$$

当且仅当 $x \in X_i, (i = 1, 2, \dots, k)$, 有 $f(x, a_R) = i$ 。这样,所有涉及知识库的定义都可用信息系统的定义来描述。

1.2 知识约简与核

随着计算机存储技术的发展,知识表达系统中的知识也呈海量级增长,在海量级的知识中存在大量重复的知识,也存在大量无关紧要的知识。从大量的知识中删除重复和无关紧要的知识的过程就是知识约简。约简与核是利用粗糙集进行知识约简时用到的两个主要概念^[7,8]。

设 $Q \subseteq P$, 如果 Q 是独立的,且 $\text{ind}(Q) = \text{ind}(P)$, 则称 Q 为 P 的一个约简,这里 P 可以有多种约简, P 的约简记为 $\text{red}(P)$ 。

P 中所有必要的属性组成的集合称为 P 的核,记作 $\text{core}(P)$ 。

核与约简有如下关系:

$$\text{core}(P) = \bigcap \text{red}(P) \quad (8)$$

其中 $\text{red}(P)$ 表示 P 的所有约简。

设 $A, Q \subseteq P, p \subseteq A$, 如果满足:

$$\text{pos}_{\text{ind}(A)}(\text{ind}(Q)) = \text{pos}_{\text{ind}(A - \{a\})}(\text{ind}(Q)) \quad (9)$$

则称 p 为 A 中 Q 不必要的;反之 p 为 A 中 Q 必要的。如果 A 中的每个属性 p 都为 Q 必要的,则称 A 为 Q 独立的。

设 $S \subseteq A, S$ 为 A 的 Q 约简,当且仅当 S 是 A 相对于 Q 独立的属性子集且满足:

$$\text{pos}_S(Q) = \text{pos}_A(Q) \quad (10)$$

A 的 Q 约简称为相对约简。 A 中所有 Q 必要的原始关系构成的集合称为 A 的 Q 核,简称为相对核,记为 $\text{core}_Q(A)$ 。

相对核与相对约简有如下关系:

$$\text{core}_Q(A) = \bigcap \text{red}_Q(A) \quad (11)$$

上式中, $\text{red}_Q(A)$ 是所有 A 的 Q 约简构成的集合。

2 知识获取

知识获取就是把用于求解专门领域问题的知识从拥有这些知识的知识源中抽取出来,并转换为特定的计算机表示。知识源包括人类专家、教科书、数据库及人本身的经验。计算机表示有产生式表示、面向对象表示、框架表示和自定义表示等。知识获取是构建专家系统花费时间最长、最困难的部分,是开发专家系统的“瓶颈”。

2.1 知识获取方式

知识获取方式在一定程度上决定着获取的知识是否准确和全面,从而影响专家系统诊断的准确性和有效性。知识获取方式多种多样,但无论哪一种方式都必须根据专家系统自身的特点并结合计算机技术。

从知识源获取知识,一般有三种方式:人工获取、自动获取和非自动获取^[9]。自动获取是指系统具有自动学习功能,自动完成知识获取;非自动获取是指获取的全部或部分工作由人工完成,它又分为两种方式:一种是通过知识工程师将领域知识转化为系统表示方式,另一种是通过智能编辑器以人机交互方式从领域专家那里获取知识,前一种称为人工获取方式,后一种称为半自动获取方式。迄今为止,在开发实用的专家系统时,还没有一个通用而有效的知识获取方法。

根据作者多年的经验,并结合知识源的类型等,确定如下几种方法^[10]:

(1) 查询图书资料及相关文档。无论过去还是现在,书籍是人们获取知识最主要的途径;对于研究的专家系统,总有与其相关的大量文档,这些文档是前人的知识和经验的积累。知识工程师通过查阅大量的图书资料及相关文档,整理出知识输入到专家系统中是获取知识最重要的方法。

(2) 与领域专家交谈。在交谈之前,知识工程师应对欲开发的专家系统有一定深度的了解,并掌握一些基本概念、过程及分析方法,而领域专家也应尽量掌握专家系统的有关原理和知识表示方式。针对欲开发专家系统的相关问题进行详细交流,反复磋商和讨论,最后将专家的知识以显化的“知识权一对象体”的形式表示成文本。

(3) 案例分析。案例分析是针对领域专家一般善于讨论具体实例而不讨论抽象术语,通过具体实例来获取知识的一种方式。由于领域专家有多年的经验积累,对欲开发的专家系统有较深入的了解,比较适合采用这种方法。优点是:知识条理清晰,容易结构化处理,便于领域专家分析讨论;有利于创建实例库,供推理时使用。

(4) Internet 网络。从网上获取知识、获取信息已成为当今重要的手段。优点是:多个不同领域专家可同时参与、领域专家可以和最终用户交流,使用方便、节省时间和费用。主要方式是:直接从网页上查找所需的知识,操作简单,但花费时间较多;通过 Email 邮件、视频会议等方式直接同领域专家进行交流、讨论,优点是快速、准确、方便。

2.2 知识获取过程

具体获取可分为三个过程^[11]。

(1) 概念化,确定专家系统有关的概念、信息、数

据等,哪些是已知的,哪些是推理得到的;

(2) 形式化,将概念、信息和数据转换成专家系统所要求的知识表示形式,建立专家系统求解模型框架,并确立推理规则和控制策略等;

(3) 知识库求精,解决知识库中存在的矛盾、错误和冗余,对知识库进行测试、检查,发现问题并进行修改,直到得到满意的结果为止。

知识获取过程一般包括两个阶段:从知识源中提取知识,并转换为系统表示形式;将这种具有系统表示形式的知识转换为具有知识库表示形式的知识,并构成知识库。知识获取过程图如图 1 所示^[10]。在知识的抽取转换过程中,属性约简是其中的重要环节。

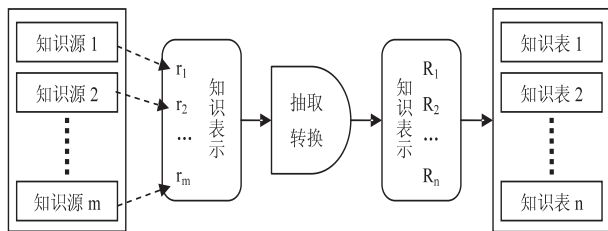


图 1 知识获取示意图

3 属性值约简算法

3.1 约简算法

对于信息系统中任意一个实例对象 X , 可得一集合族 $F = \{[x]_{a_1}, [x]_{a_2}, \dots, [x]_{a_n}\}$, 其中 $C = \{a_1, a_2, \dots, a_n\}$, $Y = [x]_D$, 且有 $\cap F \subseteq Y$, 根据集合族的概念, 可得属性值简化的概念。对于实例 X , 其属性值的简化为 F 的 Y 简化集合族中所对应的属性组成的集合。即, 若 $C' = \{a'_1, a'_2, \dots, a'_n\}$, $C' \subseteq C$, 且 $F' = \{[x]_{a'_1}, [x]_{a'_2}, \dots, [x]_{a'_n}\}$ 为 F 的 Y 简化, 即 C' 为实例 X 的属性值简化, 它代表了每个实例的无冗余的完备信息。实例 X 的属性值的核值是 F 的 Y 核集合族中所对应的属性组成的集合, 是所有属性值简化的交集^[12-14]。

定义 1: 如果对于每个 $y \neq x (x, y \in U)$, $P \subseteq C$, $d_y \mid P = d_x \mid P$, 意味着 $d_y \mid D = d_x \mid D$, 则由属性 R 下属性值可做出正确决策; 相反, 如果 $d_y \mid D \neq d_x \mid D$, 则称在属性 R 下决策规则产生冲突。

定义 2: 若删除某条决策规则 d_x 中的条件属性 a , 该条决策规则和其它决策规则产生冲突, 则称该属性 a 的属性值 $d_x(a)$ 为关键值。即 $P \subseteq C$, $d_y \mid P - \{a\} = d_x \mid P - \{a\}$ 时, $d_y \mid D \neq d_x \mid D$, 则称属性 a 的属性值 $d_x(a)$ 为关键值, 记为 $d_x^k(a)$, 对于 $\forall d_x^k(a) = R$, 则称 R 为决策规则 d_x 的核值。

3.2 约简步骤

值约简算法步骤为^[15]:

第 1 步: 逐列考察条件属性。删除某条记录的某

个属性后有下列三种情况:

- ①产生冲突,表示不可约简;
- ②未产生冲突但含有重复记录,则属性值标为“*”,表示可以约简;
- ③既不冲突又不产生重复记录,则将属性值标为“?”,表示是否可以约简待定。

第2步:删除可能产生的重复记录。若某个记录的所有条件属性均被标记,则将标有“?”的属性值修改为原值。

第3步:考察含有标记“?”的记录。有下列两种情况:

- ①仅由未被标记的属性值就可决策,则将“?”标记为“*”;
- ②仅由未被标记的属性值不可决策,则将“?”修改为原来的属性值。

第4步:删除所有属性均被标记为“*”的记录及可能产生的重复记录。

第5步:如果两条记录决策相同,删除含有“*”较少的记录。

4 约简实例

城市物流中心选址是一个涉及诸多影响因素的综合决策问题,在选址的过程中各因素都有不同程度的影响,只有将各影响因素集成起来考虑,才能使城市物流中心的选址决策更合理、更具科学性。通常考虑中心建设的三个最重要的因素指标,分别是经济因素、基础设施、自然环境,每个因素指标及选址评价分别用好、中、差来衡量。通过调查问卷及专家意见,结果如表1所示。

表1 物流中心选址决策表

对象序号	经济因素	基础设施	自然环境	选址评价
t_1	差	中	差	差
t_2	好	中	中	中
t_3	中	好	好	中
t_4	差	好	中	好
t_5	中	中	好	差
t_6	好	中	中	中
t_7	中	好	好	好
t_8	好	好	中	好

对表1中的数据,对象用 U 表示;经济因素、基础设施、自然环境,分别用 a_1 、 a_2 、 a_3 表示;好、中、差三个等级,分别用数字1、2、3表示;选址评价用 d 表示,评价结果好、中、差分别用1、2、3表示。整理后的数据如表2所示。

第1步:以记录 t_1 为例。如果删除属性 a_1 ,则 t_1 与 t_6 冲突,则 (t_1, a_1) 保留原值;如果删除属性 a_2 ,既不冲突也不重复, (t_1, a_2) 标记为“?”;如果删除属性 a_3 ,既不冲突也不重复,则 (t_1, a_3) 标记为“?”。用同样

的方法处理完 t_2 、 t_3 、 t_4 、 t_5 、 t_6 、 t_7 、 t_8 后,可以得到表3所示的决策表。

表2 数据整理后的决策表

U	a_1	a_2	a_3	d
t_1	3	2	2	1
t_2	1	1	2	2
t_3	2	3	1	2
t_4	3	1	1	3
t_5	2	2	2	1
t_6	1	2	2	2
t_7	2	1	1	3
t_8	1	1	1	3

表3 第1步值约简后的决策表

U	a_1	a_2	a_3	d
t_1	3	?	?	1
t_2	?	*	2	2
t_3	?	3	?	2
t_4	*	?	?	3
t_5	2	?	?	1
t_6	1	*	?	2
t_7	*	?	?	3
t_8	*	?	1	3

第2步:对于记录 t_4 和 t_7 ,所有条件属性均被标记,将“?”的属性值修改为原值。记录 t_4 和 t_7 重复,删除 t_7 。结果如表4所示。

表4 第2步值约简后的决策表

U	a_1	a_2	a_3	d
t_1	3	?	?	1
t_2	?	*	2	2
t_3	?	3	?	2
t_4	*	?	?	3
t_5	2	?	?	1
t_6	1	*	?	2
t_8	*	?	1	3

第3步:对于记录 t_1 ,由 $a_1 = 3$ 即可判断出决策,将 a_2 、 a_3 标记为“*”;同理,对记录 t_2 的 a_1 、对于记录 t_3 的 a_1 和 a_3 、对于记录 t_8 的 a_3 标记为“*”。对于 t_5 和 t_6 ,未标记的值不能做出决策,故将“?”修改为原来的属性值。结果如表5所示。

表5 第3步值约简后的决策表

U	a_1	a_2	a_3	d
t_1	3	*	*	1
t_2	*	*	2	2
t_3	*	3	*	2
t_4	*	1	1	3
t_5	2	2	2	1
t_6	1	*	2	2
t_8	*	*	1	3

第4步:无所有属性均被标记为“*”的记录。

第5步:对于记录 t_2 和 t_6 ,仅有一个属性值不同,且记录 t_6 可以 t_2 的规则判断,故删除 t_6 。同理,可以删除 t_4 。结果如表6所示。

5 结束语

知识获取和表示就是把解决特定问题所用的专门

表 6 第 5 步值约简后的决策表

U	a_1	a_2	a_3	d
t_1	3	*	*	1
t_2	*	*	2	2
t_3	*	3	*	2
t_5	2	2	2	1
t_8	*	*	1	3

知识从某些知识来源中提炼出来,并表示成计算机能接受和使用的方式。在专家系统中,提供准确的推理知识是进行决策规划的关键。知识获取是构造专家系统的“瓶颈”问题,专家知识的好坏直接影响整个系统的性能。

粗糙集理论作为一种处理不完备、不精确及不确定数据的有效方法,在知识获取领域发挥了重要作用并具有广泛的应用前景。

文中开展基于粗糙集的属性值约简算法研究具有重要的理论意义和现实意义,但仍然存在很多问题,比如对动态数据支持不够、处理效率与数据量成反比以及得到的规则冗余度较高等问题,这些问题需要进一步深入研究。

参考文献:

[1] 裴小兵. 粗糙集的知识约简研究[D]. 武汉: 华中科技大学, 2006.

[2] 吴守领, 杨颖, 杨磊, 等. 基于粗糙集的决策表属性约简方法的研究[J]. 计算机技术与发展, 2012, 22(1): 32-35.

[3] Guan X, Yi X, He Y. Knowledge reduction and its applications based on rough set[J]. Journal of Intelligence, 2009, 23(3): 464-467.

(上接第 153 页)

别是传统方法不能预报的时候。由于滤波对飞机位置的更准确定位和预测,能应用于进一步缩小安全距离的飞行情况,对飞行流量的增加和自由飞行的开放有很大帮助。

参考文献:

[1] 李俊菊, 宋万忠, 梁海军, 等. 中期冲突探测算法的研究与设计[J]. 计算机工程与设计, 2010, 31(20): 4492-4493.

[2] 罗世谦, 冯子亮. 一种高效的中期冲突探测随机化算法[J]. 计算机应用与软件, 2010, 27(3): 56-57.

[3] Jiang Bin, Chowdhury F. Fault estimation and accommodation for linear MIMO discrete-time systems[J]. IEEE Trans. on Control Systems Technology, 2005, 13(3): 493-499.

[4] 魏光兴, 杨昌其. 飞行冲突检测与调配的方法研究[J]. 中国民航学院学报, 2005, 23(6): 1-4.

[5] 吴舜歆, 彭炜, 李瑞芳. 飞行计划冲突探测算法研究[J].

[4] 张春燕. 基于粗糙集理论的属性值约简算法研究[J]. 计算机与现代化, 2008, 24(7): 79-81.

[5] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2003.

[6] 郑梦泽. 基于粗糙集理论的交通控制知识获取与决策研究[D]. 上海: 复旦大学, 2008.

[7] 张晓艳. 基于粗糙集的旋转机械故障诊断知识获取方法研究[D]. 苏州: 苏州大学, 2009.

[8] 代广珍, 徐超. 基于 RS 理论的快速属性约简求核方法[J]. 计算机技术与发展, 2011, 21(4): 133-135.

[9] 刘宝康, 杜玉娥. 基于“3S”的甘肃省苜蓿病害诊断专家系统的知识获取与表示[J]. 草业科学, 2008, 25(11): 88-93.

[10] 任永昌. 软件成本估算及其专家系统研究[D]. 阜新: 辽宁工程技术大学, 2008.

[11] Ren Y C, Xing T, Zhu P. An attributes reduction algorithms of expert system knowledge acquisition[J]. Applied Mechanics and Materials, 2010, 48(6): 187-191.

[12] 汪凌, 胡培. 基于粗糙集的决策系统知识获取算法及实证分析[J]. 情报杂志, 2009, 28(3): 144-147.

[13] Dai Jianhua, Li Yuanxiang, Liu Qun. A hybrid genetic algorithm for reduct of attributes in decision system based on rough set theory[J]. Wuhan University Journal of Natural Sciences, 2002, 7(3): 285-289.

[14] Zhou J, Miao D Q, Pedrycz W, et al. Analysis of alternative objective functions for attribute reduction in complete decision tables[J]. Soft Computing - A Fusion of Foundations, Methodologies and Applications, 2010, 15(8): 1601-1616.

[15] 常晓艳. 粗糙集知识约简算法研究与应用[D]. 北京: 北京化工大学, 2005.

计算机工程与设计, 2007, 27(3): 430-432.

[6] 国务院中央军委空中交通管制委员会. 飞行间隔规定[M]. 北京: 中国民航出版社, 2007.

[7] Paielli R A, Erzberger H. Conflict probability estimation for free flight[J]. Journal of Guidance, Control and Dynamics, 1997, 20(3): 588-596.

[8] Erzberger H, Paielli R A, Douglas R, et al. Conflict detection and resolution in the presence of prediction error[C]//Proc. of the 1st USA/Europe Air Traffic Management R&D Seminar. [s.l.]: [s.n.], 1997.

[9] Prandini M, Watkins O J. Probabilistic aircraft conflict detection[R]. [s.l.]: [s.n.], 2005.

[10] 梅永兵, 朱允民. 基于 Kalman 滤波的飞行冲突探测[J]. 四川大学学报(自然科学版), 2005, 42(3): 451-452.

[11] 高扬, 徐浩军, 郑海峰. 计算航路上飞行冲突概率的一种方法[J]. 飞行力学, 2009, 27(2): 50-53.

一种专家系统知识获取时的属性值约简算法

作者: [彭霞](#), [朱萍](#), [任永昌](#)
作者单位: [渤海大学 信息科学与技术学院, 辽宁 锦州121013](#)
刊名: [计算机技术与发展](#)
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2013(4)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201304040.aspx