

粗糙集理论分析及其在农业数据中的应用

郭红波¹, 周明全^{1,2}, 耿国华¹, 王小凤¹

(1. 西北大学 信息科学与技术学院, 陕西 西安 710069;

2. 北京师范大学 信息学院, 北京 100875)

摘要:从数据库中挖掘有用信息,将难理解的纯数据变为容易利用的规则,从而为以后的决策提供依据。以粗糙集理论和规则提取算法为基础,将基于信息量的粗糙集属性约简算法和规则提取算法集成起来提出一种集成算法,应用粗糙集约简掉冗余属性,然后利用规则提取算法得出有效规则。将此集成算法应用于农业领域,得出规则,并且效果良好,理论分析和应用都表明了本算法的有效性和实用性。此集成算法可以应用于各种大型数据库中,从中得出有效规则,让历史数据为以后的决策服务。

关键词:粗糙集;信息量;属性约简;值约简

中图分类号:TP391.4

文献标识码:A

文章编号:1673-629X(2007)10-0165-03

Rough Set Theory Analysis and Its Application in Agricultural Field

GUO Hong-bo¹, ZHOU Ming-quan^{1,2}, GENG Guo-hua¹, WANG Xiao-feng¹

(1. School of Information Science and Technology in Northwest University, Xi'an 710069, China;

2. School of Information, Beijing Normal University, Beijing 100875, China)

Abstract: To dig useful information from database, and get the easy using rules from difficult understand pure data, then offer future decision information according to the rules obtained in the system. Based on the rough set theory and rules extraction algorithm, an integrated algorithm is proposed. Using rough set theory the redundant attribute can be reduced, and then the efficient rules are obtained according to the rules extraction algorithm. The data in agricultural fields is analyzed by this algorithm and effective rules are obtained. The analysis in theory and the applications show that it is not only effective but also feasible. The algorithm can be widely used in all kinds of large databases, from which the effective rules can be obtained. Then the history's data can serve the future decision-making.

Key words: rough set theory; information quantity; reduction of attribute; value reduction

0 引言

粗糙集(Rough Set)理论是由 Z. Pawlak 于 1982 年提出的^[1],它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;它的主要思想是在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。它的核心内容是属性约简和规则提取。属性约简是根据属性的重要程度约简的,在属性重要性的度量

上可以采用基于属性依赖度、信息熵、互信息、加权平均和信息量度量等^[2~8]方法,在比较了上述几种方法后采用基于信息量度量^[8]的方法;在规则提取过程中,必须先进行值约简,然后导出规则。文中将用此算法分析农业数据,得到有效的规则。

1 粗糙集理论基本概念

在粗糙集理论中,对象的知识是通过指定对象的基本特征(属性)和它们的特征值(属性值)来描述的。一个知识表达系统定义^[9]为

$$S = \langle U, C, D, V, F \rangle$$

式中, U 是对象的集合, $C \cup D = A$ 是属性集合(等价关系集合),子集 C 和 D 分别称为条件属性和结果属性, $V = \bigcup_{a \in A} V_a$ 是属性值的集合, V_a 表示了属性 $a \in A$ 的范围, $f: U \rightarrow V_a$ 为一单射,使论域 U 中任一元素取属性 a 在 V_a 中的某一个唯一值。这种定义方式使对

收稿日期:2006-12-24

基金项目:国家自然科学基金资助项目(60372072);中国博士后科学基金资助项目(2003033519)

作者简介:郭红波(1978-),男,陕西人,硕士研究生,研究方向为智能挖掘和音频处理等;周明全,教授,博士生导师,研究方向为图形图像处理、数据挖掘等;耿国华,教授,博士生导师,研究方向为智能信息处理、数据库、数据挖掘等。

象的知识可以方便地以数据表格形式描述,这种数据表称为知识表达系统。

在知识表达系统^[9]中,如果属性集合 A 由条件属性集合 C 和结论属性集合 D 组成, C, D 满足 $C \cup D = A, C \cap D = \emptyset$, 则称 S 为决策系统。为了表示简单,有时用 $(U, C \cup \{d\})$ 表示决策系统,即结论属性集合只包含一个元素。

在一个决策系统中,各个条件属性之间往往存在着某些程度上的依赖或关联,简约可以理解为在不丢失信息的前提下,可以最简单地表示决策系统的结论属性对条件属性的集合的依赖和关联。通过一组相对简约,可以得到决策系统 $S = (U, C \cup \{d\})$ 中最简单的规则集,其中,每个相对简约就是一条规则的前件。所有简约的交集即为核。属性的核就是条件属性集中对于分类最为重要的那些属性,少了它们,分类的质量就明显下降^[9]。本集成算法中的 step1 ~ step3 为属性简约算法,它是采用文献[8]中的算法,算法运行的结果为简约后的属性集。算法中所用到的公式^[8]如下:

定义 1 设 $S = \langle U, A, V, f \rangle$ 是一个信息系统^[10], $P \subseteq A, U/\text{IND}(P) = \{X_1, X_2, X_3, \dots, X_n\}$, 知识 P 的信息量定义为 $I(P) = \sum_{i=1}^n \frac{|X_i|}{|U|} (1 - \frac{|X_i|}{|U|}) = 1 - \frac{1}{|U|^2} \sum_{i=1}^n |X_i|^2$, 其中 $|X|$ 表示集合 X 的基数, $|X_i|/|U|$ 表示等价类 X_i 在 U 中的概率。其中 $U/\text{IND}(P)$ 是根据属性 P 划分的等价类集合。

定义 2 设 $S = \langle U, A, V, f \rangle$ 是一个决策表, $A = C \cup D, P \subseteq A$, 知识 $Q(U/\text{IND}(Q) = \{Y_1, Y_2, \dots, Y_m\})$ (属性集合) 相对于知识 $P(U/\text{IND}(P) = \{X_1, X_2, X_3, \dots, X_n\})$ (属性集合) 的条件信息量 $I(Q|P)$ 定义为

$$I(Q/P) = \sum_{i=1}^n \frac{|X_i|}{|U|} \sum_{j=1}^m \frac{|X_i \cap Y_j|}{|X_i|} (1 - \frac{|X_i \cap Y_j|}{|X_i|})$$

定义 3 设 $S = \langle U, A, V, f \rangle$ 是一个决策表, $A = C \cup D, C$ 为条件属性, $D = \{d\}$ 为决策属性集, $R \subseteq C, a \in R$ 在 R 中的重要性 $\text{Sig}(a, R, D)$ 定义为 $\text{Sig}(a, R, D) = I(D|R \setminus \{a\}) - I(D|R)$; 当 $R = \{a\}$ 时, $\text{Sig}(a, R, D) = I(D) - I(D|\{a\})$, 其中 $U/\text{IND}(\emptyset) = \{U\}, I(D|\emptyset) = I(D)$ 。

2 决策规则

对决策表进行属性简约之后,得到一张与原决策表具有相同不可区分关系的新决策表,且决策表含有较少的属性数。通过对该表进行值简约,将删除所有

对提取规则无关的属性值。分析最小值约简,也可以从值核入手。所谓值核是指,对信息表每一条记录中的所有属性值求核,即在每条记录中寻找对得出决策影响最大的属性值。关于决策表求值核,目前国内已有一些资料对其进行了介绍^[4,11~13],本集成算法中的 step4 ~ step8 为决策规则算法^[13]。

3 集成算法描述

基于属性简约和规则提取的集成算法描述如下:

输入 决策表 $S = \langle U, A, V, f \rangle, A = C \cup D$, C 为条件属性集, D 为决策属性集。

输出 规则。

第 1 步:计算 $I(D/C)$;

第 2 步:令 $\text{Core}(C, D) = \emptyset$

for (every $a \in C$) 计算 $\text{Sig}(a, C, D)$;

if ($\text{Sig}(a, C, D) > 0$) $\text{Core}(C, D) :=$

$\text{Core}(C, D) \cup \{a\}$;

第 3 步: if ($I(D|\text{Core}(C, D)) = I(D|C)$), 则输出 $\text{Core}(C, D)$ 为 C 的最小约简, 终止; 否则将非核条件属性记入集合 Att 中, 即 $\text{Att} = C \setminus \text{Core}(C, D)$, 令 $B = \text{Core}(C, D)$,

While ($I(D|B) \neq I(D|C)$) do

{ for (every $a_i \in \text{Att}$) 计算 $I(D|B \cup \{a_i\})$

$a_j = \min \{a_i | I(D|B \cup \{a_i\})\}$

(如果有几个属性 $a_i \in \text{Att}$ 具有相同的最小信息量, 则选择属性重要性最大的属性)

$\text{Att} = \text{Att} \setminus \{a_j\}; B = B \cup \{a_j\}$; 计算新的 $I(D|B)$;

}

第 4 步: 逐条记录删除决策表中每个属性的所有值, 若决策表产生冲突记录, 则保留原属性值; 若未产生冲突且除该属性外其余属性所构成的决策表中有与本记录完全相同的记录, 则将该条记录的属性值标为“*”; 若未产生冲突且除该属性外其余属性所构成的新决策表中没有与本记录完全相同的其它记录, 则将该属性值标为“?”。

第 5 步: 删除可能产生的重复记录。

第 6 步: 检查每条含有标记“?”的记录。如果仅靠标记“?”及“*”外的其余属性即可判断出决策, 则将“?”标记为“*”, 否则修改为原属性值。如某条记录的所有条件属性均被标记为“?”和“*”, 则将标有“?”的属性项修改为原属性值。

第 7 步: 删除所有条件属性被标为 * 的记录及可能产生的重复记录。

第 8 步:如果某两条记录仅有一个条件属性对应的值不同,而且其中一条记录该属性被标为*,那么可以对该属性被标为*的记录进行分析。此记录如果仅有未被标为*的属性值即可判断出正确的决策,则删除另外一条记录,否则删除此记录。

经过约简之后得到的决策表,所有的属性值均为该表的值核,所有的记录均为该决策表的规则。

4 集成算法在农业中的应用

表 1 是一些农业数据,为了采用粗糙集,离散化后得到表 2。

表 1 原始农业数据

编号	区域	耕地面积	农业人口	农业投资	农业产出	记数
1	heilongjiang	8826.53	20084.6	343	24540	1
4	shandong	6867.87	68459.6	216	64750	1
11	ganshu	3477.13	18682.5	286	10310	1
17	jiangshu	4562.33	52707.6	192	58050	1
21	human	3318.60	51824	210	39740	1
30	xizhang	223.00	1882.24	30	1720	1

表 2 离散化后的农业数据

编号 U	区域(a1)	耕地面积(a2)	农业人口(a3)	农业投资(a4)	农业产出(d)	记数
1	东北	大	中	大	中	1
4	北	大	大	中	大	1
11	西北	中	中	中	小	1
17	东	中	大	中	大	1
21	中	中	大	中	中	1
30	西南	小	小	小	小	1

从表 2 可以看到原始属性为:{区域(a1),耕地面积(a2),农业人口(a3),农业投资(a4),农业产出(d)}。对表 2 数据运行算法第 1 步~第 3 步,得到:

- 1)“区域(a1)”的 sig 值为:0.188888892531395
- 2)“耕地面积(a2)”的 sig 值为:0
- 3)“农业人口(a3)”的 sig 值为:
0.0333333350718021
- 4)“农业投资(a4)”的 sig 值为:
0.0444444455206394

最后得到约简后的属性为:{区域(a1),农业人口(a3),农业投资(a4)},由此结果看出,“区域”是“农业产出”最重要的决定因素。其它的重要性程度依次为“农业投资,农业人口,耕地面积”。表 2 中数据提供的

信息经过算法计算得出:属性“耕地面积”可以去除。当然,此结果不一定准确,但对表 1 所提供的信息,它是完全正确的。粗糙集本身是不需要先验知识的,完全由所提供的信息得出结论。所以在运用粗糙集时一定要注意数据集的完备性。得出的结果可以给相关人员提供重要的信息,比如由此表可以得出由于各个区域的土壤不同,所以区域在很大程度上决定了农业产出。

在第 3 步的基础上,继续执行第 4 步~第 8 步,得出如下规则:

规则 1:(区域(a1),西北)得出(农业产出(d),小);

规则 2:(农业人口(a3),中)并且(农业投资(a4),大)得出(农业产出(d),中);

规则 3:(农业人口(a3),小)并且(农业投资(a4),小)得出(农业产出(d),小);

.....

这些规则相对原始数据而言,就很明了,对决策者很有用,整个算法完毕。

5 总结与讨论

在分析了多个属性约简和规则提取算法之后,挑选出基于信息量的属性约简算法和决策规则提取算法,将它们集成起来并且应用于农业数据库,可以推导出合理的规则。从这些规则,可以得出很重要的信息;最重要的是,这些规则不需要主观信息,完全是由数据库中的数据信息推导得出,具有很高的可信度。同理,此系统也可以有效地应用于其它各种领域,约掉冗余属性并且得出有效的规则。

参考文献:

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356.
- [2] 石峰, 姜臻亮, 张永清. 一种改进的粗糙集属性约简启发式算法[J]. 上海交通大学学报, 2002, 36(4): 478-481.
- [3] Qiang S, Alexios C. Amodular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems[J]. Engineering Application of Artificial Intelligence, 2000, 13(3): 263-278.
- [4] 王珏, 王任, 苗夺谦, 等. 基于 RoughSet 理论的“数据浓缩”[J]. 计算机学报, 1998, 21(5): 393-399.
- [5] Wong S K M, Ziarko W, Li Y R. Comparison of rough set and statistical methods in inductive learning[J]. Int J of Man-Machine Studies, 1986, 24: 53-73.
- [6] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机

(下转第 188 页)

2.5 系统设计的特色

本系统与传统的在线测试系统相比,系统体系结构设计比较合理,并采用当前比较先进的方法和技术实施开发,使该系统具有更高的自动化水平,也具有一定的智能性,例如出卷、改卷及知识库的维护等都是由相应的 Agent 自动完成,一定程度上减轻了教师的负担;在多人同时测试时,由于本系统各 Agent 之间使用的是一种高层 Agent 通信语言,即 KQML 通信语言,不需传递原始数据,因此减少了通信流量;扩展性得到了提高,可以容易地引入新 Agent 完成新的任务。

3 结束语

尽管 Agent 技术的理论与当前的实际应用仍存在着一定的差距,但是 Agent 技术应用到网络教学中的在线测试系统,在某种程度上,可以较好地克服原有网络教学中在线测试系统的缺陷。笔者尝试着利用面向 Agent 的开发方法,实现网络教学中的在线测试

系统,在一定程度上满足了学生和教师对在线测试系统智能性和动态性的需求。目前,在线测试是网络课程的重要组成部分,随着网络环境的不断完善、Agent 技术的逐渐成熟,基于 Agent 在线测试系统将成为未来数字教育服务领域中的重要研究课题之一。

参考文献:

[1] 伍尔德里奇.多 Agent 系统引论[M].石纯一译.北京:电子工业出版社,2003.
[2] 毛新军.面向主体的软件开发[M].北京:清华大学出版社,2005.
[3] 朱 萍,杨肖鸳.基于 Agent 技术的敏捷虚拟企业信息系統框架研究[D].昆明:昆明理工大学,2002.
[4] Agent - Oriented Software 公司. JACK™ Intelligent Agents WebBot Manual[EB/OL]. 2006 - 12 - 30. <http://www.agent-software.com/shared/resources/index.html>.
[5] 程 蕾,王泽兵.基于 Agent 技术的智能题库系统的研究与设计[D].杭州:浙江大学,2003.

(上接第 167 页)

研究与发展,1999,36(6):681-684.
[7] 陈湘晖,朱善君,吉吟东.基于熵和变精度粗糙集的规则不确定性量度[J].清华大学学报,2001,41(3):109-113.
[8] 刘振华,刘三阳,王 珏.基于信息量的一种属性约简算法[J].西安电子科技大学学报:自然科学版,2003,30(6):835-838.
[9] Pawlak Z. Rough sets, theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[10] 梁吉业,曲开社,徐宗本.信息系统的属性约简[J].系统工程理论与实践,2001,21(12):76-80.
[11] 王 珏,苗夺谦,周育健.关于 Rough Set 理论与应用的综述[J].模式识别与人工智能,1996,9(4):337-344.
[12] 曾黄麟.粗集理论及其应用——关于数据推理的新方法(修订版)[M].重庆:重庆大学出版社,1998:83-87.
[13] 常翠云,吴 渝.一种基于 Rough Set 理论的属性约简及规则提取方法[J].软件学报,1999,10(11):1206-1211.

(上接第 183 页)

表 1 采样周期与处理时间关系表

采样周期(μ s)	计算一个采样点所需平均时间(μ s)
1	16~17
2	9~10
4	5~6
6	4~5
8	2~4

4 结束语

硬件相位计性能指标不稳定,制作费用昂贵,存在使用寿命问题,软件方法节省了硬件必需的成本费和维修费,处理更加灵活方便。而且软件处理所得密度的时间分辨率取决于探测信号的采样频率,要想获得高时间分辨率的密度,提高信号采样频率即可,硬件相位计的时间分辨率则只取决于信号调制周期。近年来高响应频率的采集卡已非常普遍,计算机的处理能力

也一直在不断提高,给软件方法的应用提供了有利的条件。

参考文献:

[1] Gao Xiang, Jie Yinxian, Xia Chengyi, et al. High Density Operation on HT - 7 Superconducting tokamak[J]. Nuclear Fusion, 2000, 40(11):1875-1883.
[2] 揭银先. HT - 7 超导托卡马克上 HCN 激光干涉仪及等离子体密度行为的研究[D].合肥:中科院等离子体物理研究所,2002.
[3] Jie Y X, Gao X, Cheng Y F. Multi - channel FIR HCN Laser Interferometer on HT - 7 Tokamak[J]. International Journal of Infrared and Millimeter Waves, 2000, 21(9):1375-1380.
[4] 陈俊本,高 翔,郭其良,等.一种计算远红外干涉仪相位差的数值方法[J].核聚变与等离子体物理,1989,9(3):189-192.
[5] 张春飞,罗家融.软件去除零点漂移方法的讨论[J].计算机测量与控制,2004,12(7):684-686.