

应用粗糙模糊度的不完备信息系统属性约简

汪琼枝^{1,2}, 毛军军^{1,2}, 吴涛^{1,2,3}, 李萍^{1,2}

(1. 安徽大学 智能计算与信号处理教育部重点实验室, 安徽 合肥 230039;

2. 安徽大学 数学科学学院, 安徽 合肥 230039;

3. 南京大学 计算机软件新技术国家重点实验室, 江苏 南京 210093)

摘要:粗糙集理论能有效地处理不精确、不一致、不完整等不完全数据信息,可以对数据信息进行分析和推理,发掘隐含知识,揭示潜在规律。属性约简是粗糙集理论的重要研究课题。在现实生活中,由于各种条件限制,信息的不完备现象广泛存在,限制了经典 Rough 集理论在一些实际问题中的应用。文中引入粗糙模糊度度量,定义了一种新的知识熵。在此基础上,提出了一种基于信息观下粗糙模糊度的不完备信息系统属性约简算法。通过仿真实验说明了该算法的有效性和较好的时间优越性。

关键词:粗糙模糊度;不完备信息系统;属性约简

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2010)10-0001-03

An Attribute Reduction Algorithm for Incomplete Information Systems Based on Rough Fuzzy Degree

WANG Qiong-zhi^{1,2}, MAO Jun-jun^{1,2}, WU Tao^{1,2,3}, LI Ping^{1,2}

(1. Ministry of Education Key Laboratory of Intelligent Computing & Signal Processing,

Anhui University, Hefei 230039, China;

2. School of Mathematical Sciences, Anhui University, Hefei 230039, China;

3. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

Abstract: Rough set theory can effectively deal with incomplete data which is imprecise or inconsistent. It can analyze the data to dig implicit information and reveal the potential law. In rough set theory, attribute reduction is an important research subject. The widespread presence of the incomplete information system limit the application of classical rough set theory. Introduces a new entropy based on the rough fuzzy metrics. On this basis, an attribute reduction algorithm based on rough fuzzy degree for incomplete information systems is proposed. The simulation experiment illustrates the effectiveness of algorithm and a better time for superiority.

Key words: rough fuzzy metrics; incomplete information systems; attribute reduction

0 引言

Rough 集作为一种刻画不完整性和不确定性的数学理论,能有效地处理不精确、不一致、不完整等各种不完全数据信息,还可以对数据进行分析和推理,发掘隐含的知识,揭示潜在的规律。粗糙集理论最初由波兰数学家 Z. Pawlak 于 1982 年提出,由于初期的研究

大多用波兰语发表的,因此并没有引起国际数学界和计算机界学者的关注,到 20 世纪 90 年代初已引起各国学者的关注,1992 年,在波兰召开第一届关于粗糙集理论的国际学术会议。1995 年,ACM Communication 将粗糙集理论列为计算机科学新浮现研究课题。1998 年,Information Sciences 为粗糙集理论研究出了一期专辑。中国有关粗糙集理论研究的文章大多出现在 1997 年之后,近年来趋热。粗糙集理论目前被广泛应用于模式识别、机器学习、数据挖掘、知识获取,决策分析和智能控制等许多领域。属性约简是粗糙集理论中的一个重要的研究课题,其主要思想是:在保持信息系统分类能力不变的前提下,找到原始数据属性集的一个最小子集,通过属性约简,达到发掘知识并简化知识的目的。约简的概念相当于机器学习中的属性子集选

收稿日期:2010-01-30;修回日期:2010-05-18

基金项目:中国博士后基金面上项目(20070411028);国家自然科学基金(60675031);安徽省自然科学基金项目(KJ2008B093)

作者简介:汪琼枝(1983-),女,安徽六安人,硕士研究生,研究方向为智能计算理论与应用;毛军军,博士,副教授,研究方向为粒计算及其应用;吴涛,博士,教授,研究方向为机器学习、智能计算及其应用。

择问题,但是约简的数学意义更简洁明确。约简结果通常不唯一。许多学者从不同的角度提出了获取信息系统和决策系统属性的约简算法^[1,2]。这些算法大体上可以分为三大类:基于代数理论的^[3],基于区分矩阵和区分函数构造的^[4-7],基于信息熵理论的^[8,9]。对于不同的数据和应用,这些算法的效果是不同的,各有优缺点。这些算法极大地丰富了粗糙集理论,广泛应用于税务、商业、医院等许多领域,并为其进一步发展和应用打下了坚实的基础。

经典 Rough 集理论以完备信息系统为研究对象,文献[10,11]中对完备信息系统属性约简做了讨论。然而,在自然科学、社会科学和工程技术的很多领域中,都不同程度地涉及到对不确定因素和不完备信息的处理,由于对原始数据测量提取的过程中产生了误差,以及原始数据在理解或获取过程中的限制等种种原因,使得得到的原始数据是不完备信息系统。如果对这些不完备信息系统进行粗略的简单处理,效果往往不理想,甚至完全背离事实;相反,如果能够对这些不完备信息进行恰当的处理,常常有助于很多实际问题的快速高效解决。多年来,经过各国数学家、逻辑学家和计算机研究人员的努力,提出了很多基于不完备信息系统的新的粗糙集理论,这些理论能够有效地表达不确定、不精确或模糊的知识,并利用这些不完备的信息推理判断并解决许多实际问题。文献[12]中提出了基于信息熵的不完备信息系统属性约简算法 IEAR-A 算法,该算法引入的信息熵反映了属于集合 X 的元素“贡献”的不确定性。文中应用粗糙模糊度度量,同时考虑属于和不属于集合 X 的元素“贡献”的不确定性,定义了一种新的知识熵。提出了一种应用粗糙模糊度度量的不完备信息系统属性约简算法。最后通过仿真实验说明该算法对于不完备信息系统的属性约简不仅是可行的,而且相对于文献[12]具有较好的时间优越性。

1 相关基本概念

“知识”,在不同的范畴有不同的含义,在粗糙集理论中,“知识”被认为是一种基于对现实或抽象对象的分类能力。

信息系统用 $S = (U, A, V, f)$ 来表示,其中 $U = \{x_1, x_2, \dots, x_n\}$ 是论域,是一个非空的有限讨论对象集; $A = \{a_1, a_2, \dots, a_m\}$ 是属性集; $V = \bigcup_{a \in A} V_a$ 是值域, V_a 是属性 $a \in A$ 的值域; f 是信息函数: $U \times A \rightarrow V$, $f_a(x)$ 表示 x 在 a 的取值, $f_a(x) = (a, x) \in V_a$ 。

信息系统可以用一般表来表示,称为信息表。信息

表中每一个属性之间是等价关系,若 $A = C \cup D$, C 是条件属性, D 是决策属性,此时信息表称为决策表。若信息系统中无缺失数据,则称为完备信息系统,反之,如果信息系统中存在缺失数据,则称为是一个不完备信息系统。

对于 $X \subset U$, R 是 U 上的一个等价关系, $U/R = \{x_1, x_2, \dots, x_k\}$, 若存在 $x_{i_1}, x_{i_2}, \dots, x_{i_j} \in U/R$ 使得 $X = \bigcup_{i=1}^j x_{i_j}$, 则称 X 是关于 R 的精确集, 否则称 X 是关于 R 的粗糙集。

记:

$\underline{R}X = \bigcup \{Y \in U/R, Y \subset X\} = \bigcup \{[x]_R \mid [x]_R \subset X, x \in U\}$ 为 X 关于 R 的下近似;

$\overline{R}X = \bigcup \{Y \in U/R, Y \cap X \neq \emptyset\} = \bigcup \{[x]_R \mid [x]_R \cap X \neq \emptyset, x \in U\}$ 为 X 关于 R 的上近似。

记: $\text{ind}(P) = \bigcap_{R \in P} R = \bigcap P$ 表示 P 中所有等价关系的交集。

设 Ω 是一个等价关系簇, $R \in \Omega$, 如果 $\text{ind}(\Omega - \{R\}) = \text{ind}(\Omega)$ 则称 R 在等价关系簇 Ω 中是不必要的, 否则称 R 在等价关系簇 Ω 中是必要的。若 Ω 中每个等价关系 R 都是必要的, 则称 Ω 是独立的。

定义 1. 若 $Q \subset P$ 满足: (1) Q 是独立的; (2) $\text{ind}(Q) = \text{ind}(P)$ 则称 Q 是 P 的一个约简, 记为 $Q \in \text{Redu}(P)$, $\text{Redu}(P)$ 称为 P 的约简全体。

定义 2. 给定信息系统 $S = (U, C, V, f)$, 对于具有遗漏属性值的属性子集 $B \subseteq C$, 记遗漏值为 “*”, 容差关系 T 的定义为:

$$T = \{(x, y) \mid x \in U \wedge y \in U \wedge c_i \in B \Rightarrow (c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *)\}$$

定义 3. 在给定的二元关系 R 下, B 是一个属性子集, 记:

$$U/\text{SIM}(B) = \{R_B(x) \mid x \in U\}$$

式中, $R_B(x) = \{y \in U \mid R(x, y)\}$ 是对象 $x \in U$ 关于属性集 B 的相似类。

定义 4. 设决策表 $S = (U, C \cup \{d\}, V, f)$, 其中 U 为一个非空的有限对象集, C 是条件属性集, $\{d\}$ 是决策属性集, 对每个属性 $c \in C \cup \{d\}$, 定义信息函数 $f_c: U \rightarrow V$ 。属性集合 $P \subseteq C$, 对于任意属性 $c \in C/P$ 的重要性 $\text{SGF}(c, P)$ 定义为:

$$\text{SGF}(c, P) = H(P) - H(P \cup \{c\})$$

式中, $\text{SGF}(c, P)$ 的值越大, 说明属性 c 对属性集合 P 的分类能力影响越大, 即对于属性集合 P 越重要。

定义 5. 设决策表系统为 $S = (U, C \cup \{d\}, V, f)$, 条件属性集 C 的熵为 $H(C)$ 。称属性集 $P \subseteq C$ 是决策表系统 S 的一个熵约简, 当且仅当 $H(P) \leq H(C)$,

且对于任意的属性子集 $P' \subset P$ 都有 $H(P') \succ H(C)$ 。

2 应用粗糙模糊度度量的不完备信息系统的属性约简

定义 6. 设决策表系统为 $S = (U, C \cup \{d\}, V, f)$, 其中 U 为一个非空的有限对象集, C 是条件属性集, $\{d\}$ 是决策属性集, 决策属性 d 的值域为 $V_d = \{v_1^d, v_2^d, \dots, v_m^d\}$, 则对于集合 $X \subseteq U$, 定义其基于信息观下粗糙集的模糊度量:

$$d_z(F_X) = -\frac{1}{|U|} \sum_{j=1}^m p_j \ln p_j + (1 - p_j) \ln(1 - p_j) \quad (1)$$

式中, $p_j = k_j / |X|$, k_j 为集合 X 中决策属性值为 v_j^d 的实例个数, $|X|$ 为集合 X 的基数。

定义 7. 设决策表系统为 $S = (U, C \cup \{d\}, V, f)$, 其中 U 为一个非空的有限对象集, C 是条件属性集, $\{d\}$ 是决策属性集, 对每个属性 $c \in C \cup \{d\}$, 定义信息函数 $f_c: U \rightarrow V_c$, $V = \bigcup_{c \in C \cup \{d\}} V_c$, R 为给定的二元关系, 属性集合 $P \subseteq C$, $U/\text{SIM}(P) = \{R_p(x) \mid x \in U\}$ 。定义的模糊熵为:

$$H(P) = \frac{1}{|U|} \sum_{x \in U} d_z(F_{R_p(x)})(1 - d_z(F_{R_p(x)}))$$

易知, $H(P) \geq 0$ 。若 $H(P)$ 越大, 则说明根据知识 P 在论域 U 上分类后, $R_p(x)$ 中各决策属性值的分布越均匀, 即混乱程度越大, 说明知识 P 对信息系统 S 的分类能力越弱; 反之, 则说明知识 P 对信息系统 S 的分类能力越强。

对于相容的不完备决策表, 式(1)由 $p_j \ln p_j$ 和 $(1 - p_j) \ln(1 - p_j)$ 两部分构成, 前者主要反映属于集合 X 的元素“贡献”的不确定性, 后者主要反映不属于集合 X 的元素“贡献”的不确定性。这两部分同时考虑才能更精确地刻画粗糙集的不确定性。这就有利于以其为基础, 设计出一种基于粗糙模糊度的不完备信息系统的属性约简算法 (based on fuzzy degree of attributes reduction algorithm)。简记为: FDARA 算法。

FDARA 算法:

输入: 一个决策表系统为 $S = (U, C \cup \{d\}, V, f)$, 其中 U 为一个非空的有限对象集, C 是条件属性集, $\{d\}$ 是决策属性集, 决策属性 d 的值域为

$$V_d = \{v_1^d, v_2^d, \dots, v_m^d\}$$

输出: 该决策表系统的一个约简 Redu。

s_1 : 初始化 $\text{Redu} = \emptyset$; $\text{CAAttr} = C$; $e = H(C)$;

s_2 : 若 $H(\text{Redu}) \succ e$, 则:

(1) 根据定义 5、定义 6, 计算 $H(\text{Redu})$;

(2) 对于每个条件属性 $c_i \in \text{CAAttr}$, 计算 $H(\text{Redu} \cup \{c_i\})$;

(3) 计算每个属性 $c_i \in \text{CAAttr}$ 的重要性:

$$\text{SGF}(c_i, \text{Redu}) = H(\text{Redu}) - H(\text{Redu} \cup \{c_i\})$$

(4) 找到属性 c_j 使得 $\text{SGF}(c_j, \text{Redu}) = \text{Max}\{\text{SGF}(c_i, \text{Redu})\}$

(5) $\text{Redu} = \text{Redu} \cup \{c_j\}$, $\text{CAAttr} = \text{CAAttr} / \{c_j\}$ 并记录 Redu 中各属性加入的顺序;

s_3 : 若 $H(\text{Redu}) \leq e$, 则按加入到 Redu 的属性的顺序, 从后至前, 逐个检查每个属性 $\text{SGF}(c, \text{Redu} / \{c\}) \leq 0$, ($c \in \text{Redu}$)。若成立, 则 $\text{Redu} = \text{Redu} / \{c\}$ 。否则 $\text{Redu} = \text{Redu}$ 。

s_4 : 算法结束。

3 仿真实验

仿真实验所用计算机配置如下: CPU: AMD Athlon64(2800+), 1.0GB 内存。实验中选用了 UCI data 和 Rose 中的 6 个数据集 (见表 1), 在容差关系下对文献[12]中的 IEARA 算法和文中的 FDARA 算法进行了分析比较, 见表 2。

实验结果表明, 在容差关系下, 两种约简算法的效果相当。文中提出的 FDARA 算法在约简时间方面要优于 IEARA 算法。因此, 通过仿真实验说明了所提出的 FDARA 算法对于不完备信息系统属性约简不仅是可行的而且具有较好的时间优越性。

表 1 测试数据集

序号	数据集名称	数据集来源	样本容量	条件属性个数	完备
1	Soybean-large	UCI	307	35	否
2	Primary-tumor	UCI	339	17	否
3	Breast-cancer	UCI	286	9	否
4	Vote-isf	ROSE	300	16	是
5	URAZY-ISF	ROSE	80	22	是
6	Breast cancer wisconsin	UCI	699	9	否

表 2 基于容差关系的属性约简算法比较

数据集序号	FDARA 算法约简结果	FDARA 算法约简时间(毫秒)	IEARA 算法约简结果	IEARA 算法约简时间(毫秒)
1	12	48078	12	49500
2	16	42563	16	44110
3	8	4672	8	4750
4	8	12734	8	13000
5	7	906	7	969
6	5	25343	5	257815

4 结束语

现实生活中需要处理的数据大多是不完备的, 传统方法的采用数据补齐的方法使不完备信息系统变为

(下转第 7 页)

锻炼],并以此在男性老年组 BMI 超出 23 的节点下构成新的关联层,如此进行不断深入的分析,最终通过最小支持度和可信度的对比就可得出一套没有血脂高家族史的老年人群中预防血脂高发上的行之有效的医疗保健方案。如果将每一层的各个项称为节点,层与层之间的关系称为节点间的链接,显然多层次关联规则的挖掘过程与计算机领域内的数据结构理念如出一辙^[11]。

4 结束语

针对病历随访数据库系统中海量的医疗数据进行数据挖掘,以此实现提供疾病诊断、管理决策分析是一种发展趋势^[12]。建立符合医疗过程的数据库结构是基础,从海量的诊疗数据中挖掘数据间的关联关系则是知识性医疗数据挖掘能否正确实现的前提保证,选择快速、有效的关联规则算法是能否最终提供有价值的决策信息的关键,文中所介绍的关联规则挖掘 Apriori 算法针对医学数据所具有的多态性、不完整性、较强的时间性、复杂性、冗余性和不一致等特点的知识挖掘是成功的,能够满足病历随访数据库系统的数据挖掘需求。

参考文献:

- [1] Luo Zhiping, Zhou Xinzhi, Gu Zhongbi. Time series hydrologi-

(上接第 3 页)

完备的信息系统,很可能破坏原始数据信息系统,使得处理效果不理想,甚至出现完全与事实相悖离的重大错误。针对这类问题,在信息观下引入了粗糙模糊度量,定义了一种新的知识熵,在此基础上,提出了一种基于信息观下粗糙模糊度的不完备信息系统属性约简算法。仿真实验结果说明了该算法的有效性和较好的时间优越性。对于算法的优化及其应用,有待进一步研究讨论。

参考文献:

- [1] 覃伟荣,秦亮曦.基于粗糙集理论的条件属性动态约简算法[J].计算机技术与发展,2008,18(8):23-25.
- [2] Kryszkiewicz M, Rybinski H. Finding Reducts in Composed Information Systems. Rough Sets, Fuzzy Sets and Knowledge Discovery[C]//Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD'93). [s. l.]: [s. n.], 1994:261-273.
- [3] Xiao J M, Zhong T F. New Rough set Approach to knowledge Reduction in Decision Table[C]//proceeding of the International conference on Machine learning and Cybernetics. Shang-

cal based on data mining[J]. Computer Engineering and Applications, 2007, 43(30):231-233.

- [2] Agrawal R, Srikant R. Mining sequential patterns[C]//Proc. of the 11th Int Conf. on Data Engineering. Los Alamitos, CA: IEEE Computer Society, 1995:3-14.
- [3] Zaki M. SPADE: An efficient algorithm for mining frequent sequences[J]. Machine Learning, 2000, 42(3):31-60.
- [4] 季伟东. 一种 Apriori 算法的改进[J]. 计算机工程与科学, 2009(9):68-70.
- [5] 王 华. 医学数据挖掘中的数据预处理与 Apriori 算法改进[J]. 计算机系统应用, 2009(9):94-97.
- [6] 孙 明. 基于层次关联规则的日志本体事件领域关系学习[J]. 计算机应用研究, 2009(10):3683-3686.
- [7] 王 敏. Apriori 算法在税务系统中的应用[J]. 计算机技术与发展, 2009, 19(11):175-178.
- [8] 冯敬益. 基于 Web 的数据挖掘技术[J]. 中国科技信息, 2009(18):92-93.
- [9] 郑海波. 医疗信息的数据库管理模式[J]. 福建电脑, 2008, 24(9):172-178.
- [10] 邹长忠. 分布式数据库的关联规则更新算法[J]. 福州大学学报:自然科学版, 2008, 36(5):655-659.
- [11] 孟凡荣. 基于云理论的煤矿安全监测数据关联规则挖掘[J]. 小型微型计算机系统, 2008(9):1622-1626.
- [12] 徐 刚. 数据挖掘及其在医学领域中的应用和展望[J]. 实用临床医学, 2006(11):196-198.

hai, China: [s. n.], 2004:2208-2211.

- [4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6):681-684.
- [5] 汪小燕, 杨思春. 一种基于分辨矩阵的新的属性约简算法[J]. 计算机技术与发展, 2008, 18(2):77-79.
- [6] 汪小燕, 杨思春. 基于改进的二进制可分辨矩阵的核增量式更新方法[J]. 计算机技术与发展, 2009, 19(1):97-99.
- [7] Wang J, Wang J. Reduction Algorithms Based on Discernible Matrix: The ordered Attributes method[J]. Journal Computer Science Technology, 2001, 16(6):498-504.
- [8] 王国胤, 于 洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7):759-766.
- [9] 陈锦禾, 沈 洁. 基于信息熵的主动学习半监督分类研究[J]. 计算机技术与发展, 2010, 20(2):110-113.
- [10] 纪 滨. 信息熵在粗糙集中衍生的几个概念[J]. 计算机技术与发展, 2008, 18(6):73-75.
- [11] 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.
- [12] 付 昂, 王国胤, 胡 军. 基于信息熵的不完备信息系统属性约简算法[J]. 重庆邮电大学学报, 2008, 20(5):586-592.