

一种基于知识粒度的不完备信息系统的属性约简算法^{*})

李秀红¹ 史开泉²

(山东经济学院统计与数学学院 济南 250014)¹ (山东大学数学与系统科学学院 济南 250100)²

摘要 在不完备信息系统中,通过引入知识粒度的概念,对属性的重要性进行了定义,并以属性重要性为启发式信息,提出一种基于知识粒度的属性约简启发式算法,该算法的时间复杂度是多项式的,最后通过例子说明,该算法能得到不完备信息系统的一个约简。

关键词 不完备信息系统,知识粒度,重要度,属性约简

A Knowledge Granulation-based Algorithm for Attribute Reduction under Incomplete Information Systems

LI Xiu-Hong¹ SHI Kai-Quan²

(School of Statistics & Mathematics, Shandong Economic University, Jinan 250014)¹

(School of Mathematics & System Science, Shandong University, Jinan 250100)²

Abstract By introducing the concept of knowledge granulation, the significance of attributes is defined under incomplete information systems. With the significance of attributes as heuristic information, a heuristic algorithm based on knowledge granulation for reduction of attributes is presented, and the time complexity of this algorithm is analyzed. Finally, the experimental results show that this algorithm can find the reduct of an incomplete information system.

Keywords Incomplete information system, Knowledge granulation, Significance, Attribute reduction

1 引言

Z. Pawlak 于 1982 年提出的粗糙集(Rough Set)理论^[1]是一种处理不精确、不完全信息的软计算方法。其主要思想是:在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的分类或决策规则。近年来,该理论在计算机科学及多种领域获得了成功的应用。

知识约简是粗糙集理论的核心内容之一^[2],它刻画了信息系统中知识的本质部分。人们对基于传统信息系统的知识约简问题已进行了大量研究^[3~6]。近年来,基于不完备信息系统的知识约简已引起人们的广泛关注^[7~11]。文^[12]建立了不完备信息系统下的知识粒度理论。本文通过引入知识粒度的概念,对不完备信息系统中属性的重要性进行了定义,并以属性重要性为启发式信息,提出一种基于知识粒度的属性约简启发式算法,该算法的时间复杂度是多项式的,并通过例子说明,该算法是有效的。

2 不完备信息系统中的基本概念

定义 2.1^[7,8] 四元组 $S=(U, A, V, f)$ 是一个信息系统,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 表示属性的值域; $f: U \times A \rightarrow V$ 是信息函数,它为每个对象的每个属性赋予一个信息值,即 $\forall a \in A, x \in U, f(x, a) \in V_a$ 。

如果对于至少一个属性 $a \in A$, V_a 包含空值(用 $*$ 表示),即至少一个属性 $a \in A$, 存在一个 $x \in U$, $f(x, a)$ 未知(即 $f(x, a) = *$),则称信息系统是不完备的,否则称信息系统是完备的。这表明完备信息系统是不完备信息系统的一种特殊情况。

形。通常, $S=(U, A, V, f)$ 也称为知识表达系统或知识 A , 简记为 $S=(U, A)$ 。

定义 2.2^[7,8] 设 $S=(U, A)$ 是一个不完备信息系统, $P \subseteq A$ 是一个属性子集,由 P 确定的相容关系 $SIM(P)$ 定义为:

$SIM(P) = \{(x, y) \in U \times U \mid \forall a \in P, f(x, a) = f(y, a) \text{ 或 } f(x, a) = * \text{ 或 } f(y, a) = *\}$ 简记由属性 $a \in A$ 确定的相容关系为 $SIM(\{a\})$ 。

相容关系满足自反性和对称性,但一般不是传递的。

性质 2.1 $SIM(P) = \bigcap_{a \in P} SIM(\{a\})$ 。

令 $S_P(x) = \{y \in U \mid (x, y) \in SIM(P)\}$ 表示相对于 P 与 x 相容的所有对象的全体,称为相容类或信息颗粒,相容类一般不构成 U 的划分,它们构成 U 的覆盖。

性质 2.2 设 $P \subseteq Q \subseteq A$, 则 $S_A(x) \subseteq S_Q(x) \subseteq S_P(x)$, $x \in U$ 。

设 $U/SIM(A) = \{S_A(x) \mid x \in U\}$ 表示分类,则最细分类和最粗分类分别表示为^[12]:

$\hat{A}(U) = \{S_A(x) = \{x\} \mid x \in U\}$,

$\check{A}(U) = \{S_A(x) = U \mid x \in U\}$,

如果论域不会引起混淆,亦可将 $\hat{A}(U)$ 和 $\check{A}(U)$ 简记为 $\hat{A}$ 和 $\check{A}$ 。

定义 2.3 设 $S=(U, A)$ 是一个不完备信息系统,如果 $\forall x \in U, S_{A-\{a\}}(x) = S_A(x)$, 则称 $a \in A$ 在 A 中是不必要的或冗余的,否则,称 a 在 A 中是必要的;如果每个属性 $a \in A$ 在 A 中是必要的,则称属性集 A 是独立的,否则,称 A 是依赖的; A 中所有必要的属性组成的集合称为属性集 A 的核,记作 $Core(A)$ 。

^{*})基金项目:国家自然科学基金资助项目(70271048)。李秀红 副教授,博士生,主要研究方向:粗糙集理论与应用;史开泉 教授,博士生导师,主要研究方向:粗系统理论与应用,模糊系统理论与应用。

定义 2.4 设 $S=(U,A)$ 是一个信息系统, $P \subseteq A$, 如果 $\forall x \in U, S_P(x) = S_A(x)$, 且 P 是独立的, 则称 P 是 A 的一个约简。

性质 2.3 $\text{Core}(A) = \bigcap \text{Red}(A)$, 其中 $\text{Red}(A)$ 表示 A 的一个约简。

性质 2.3 表明, 核是所有约简的交集。

3 不完备信息系统的知识粒度与属性重要度

定义 3.1^[12] 设 $S=(U,A)$ 是一个不完备信息系统, $U/\text{SIM}(A) = \{S_A(x_1), S_A(x_2), \dots, S_A(x_{|U|})\}$, 则 A 的知识粒度定义为:

$$\text{GK}(A) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |S_A(x_i)|$$

其中 $|X|$ 表示集合 X 的基数。若 $U/\text{SIM}(A) = \hat{A}$, 则 $\text{GK}(A) = 1/|U|$; 若 $U/\text{SIM}(A) = A$, 则 $\text{GK}(A) = 1$ 。一般情况下, $1/|U| \leq \text{GK}(A) \leq 1$ 。

知识粒度可以描述知识的区分能力, 知识粒度越小, 其区分能力越强。对于知识粒度的概念, 已有以下一些结果。

定理 3.1^[12] 设 $S=(U,A)$ 是一个不完备信息系统, $P, Q \subseteq A$, 如果对 $\forall i \in \{1, 2, \dots, |U|\}$, 有 $S_P(x_i) \subseteq S_Q(x_i)$, 则 $\text{GK}(P) \leq \text{GK}(Q)$ 。

定理 3.2^[12] 设 $S=(U,A)$ 是一个不完备信息系统, $P, Q \subseteq A$, 如果对 $\forall i \in \{1, 2, \dots, |U|\}$, 有 $S_P(x_i) = S_Q(x_i)$, 则 $\text{GK}(P) = \text{GK}(Q)$ 。

利用知识粒度, 可以分析不完备信息系统中每一个属性的重要性。

定义 3.2 设 $S=(U,A)$ 是一个不完备信息系统, 属性 $a \in A$ 在 A 中的重要度定义为:

$$\text{Sig}_{A-\{a\}}(a) = \text{GK}(A - \{a\}) - \text{GK}(A)$$

定义 3.2 表明, 属性 a 在 A 中的重要性是用由 A 中去除 a 后引起的知识粒度变化的大小来度量。也就是说, 对于一个属性集合, 去掉一个属性引起的知识粒度变化量越大, 则该属性对此属性集就越重要, 反之亦然。特别地, 单个属性的重要度定义为:

$$\text{Sig}(a) = \text{Sig}_\emptyset(a) = 1 - \text{GK}(\{a\})$$

性质 3.1 $0 \leq \text{Sig}_{A-\{a\}}(a) \leq 1 - 1/|U|$ 。

性质 3.2 属性 $a \in A$ 在 A 中是必要的当且仅当 $\text{Sig}_{A-\{a\}}(a) > 0$ 。

性质 3.3 $\text{Core}(A) = \{a \in A \mid \text{Sig}_{A-\{a\}}(a) > 0\}$ 。

定义 3.3 设 $S=(U,A)$ 是一个信息系统, 任意属性 $a \in A - C$ 对于 C 的重要度定义为:

$$\text{Sig}_C(a) = \text{GK}(C) - \text{GK}(C \cup \{a\})$$

定义 3.3 表明, 属性 $a \in A - C$ 对于属性集 C 的重要度是由 C 中添加 a 后引起的知识粒度的变化大小来度量的, 也就是说, 对于一个属性集合, 添加一个属性引起的知识粒度变化量越大, 则该属性对此属性集就越重要, 反之亦然。因此可用它作为启发知识以减少搜索空间来寻找最小属性约简。

由定义 2.4, 定理 3.2 和性质 3.2, 容易得到下面的结论。

定理 3.3 设 $S=(U,A)$ 是一个不完备信息系统, $P \subseteq A$, 若 $\text{GK}(P) = \text{GK}(A)$ 且对任意 $a \in P$, 有 $\text{Sig}_{P-\{a\}}(a) > 0$, 则 P 为 A 的一个约简。

定理 3.3 实际上是从知识粒度角度提供了一种属性约简的判断方法, 我们将以此为基础, 给出一种新的基于知识粒度的属性约简算法。

4 一种基于知识粒度的不完备信息系统的属性约简算法

由性质 3.3 求出属性集 A 的核 $\text{Core}(A)$, 由于核的唯一性, 故可以将核作为求约简的起点, 并由定义 3.3 逐个选择最重要的属性增加到核中去, 直到其粒度等于整个属性集 A 的粒度为止, 这时得到的属性子集便构成系统的一个约简。下面给出求属性约简的具体算法。

不完备信息系统的属性约简算法:

输入: 一个不完备信息系统 $S=(U,A)$ 。

输出: $S=(U,A)$ 的一个属性约简。

步骤 1: 计算属性集 A 的知识粒度 $\text{GK}(A)$ 。

步骤 2: 计算属性集 A 的核 $\text{Core}(A)$ 。对 $\forall a \in A$, 计算属性 a 在 A 中的重要度 $\text{Sig}_{A-\{a\}}(a)$, 得到核 $\text{Core}(A) = \{a \in A \mid \text{Sig}_{A-\{a\}}(a) > 0\}$, 若 $\text{GK}(\text{Core}(A)) = \text{GK}(A)$, 则 $\text{Core}(A)$ 为最小约简, 输出 $\text{Core}(A)$, 算法终止, 否则转向步骤 3。

步骤 3: 令 $\text{Red}(A) = \text{Core}(A)$, 对属性集 $A - \text{Red}(A)$ 重复操作:

(1) 对 $\forall a \in A - \text{Red}(A)$, 计算属性 a 对于 $\text{Red}(A)$ 的重要度 $\text{Sig}_{\text{Red}(A)}(a)$;

(2) 选择 a , 使 $\text{Sig}_{\text{Red}(A)}(a) = \max_{a' \in A - \text{Red}(A)} \text{Sig}_{\text{Red}(A)}(a')$, 令 $\text{Red}(A) = \text{Red}(A) \cup \{a\}$;

(3) 判断 $\text{GK}(\text{Red}(A)) = \text{GK}(A)$ 是否成立, 若成立, 则 $\text{Red}(A)$ 为系统的一个约简, 输出 $\text{Red}(A)$, 算法终止, 否则, 转向(1)。

下面分析算法的时间复杂度。

计算核共需计算 $|A|$ 次 $\text{Sig}_{A-\{a\}}(a)$ 并判断其是否大于 0, 计算一次 $\text{Sig}_{A-\{a\}}(a)$ 的时间复杂度为 $O(|A| \cdot |U|^2)$, 故求核的时间复杂度 $O(|A|^2 \cdot |U|^2)$;

计算约简最多需计算 $|A| + (|A| - 1) + \dots + 1 = |A|(|A| + 1)/2 = O(|A|^2)$ 次 $\text{Sig}_{\text{Red}(A)}(a)$, 计算一次 $\text{Sig}_{\text{Red}(A)}(a)$ 的时间复杂度为 $O(|A| \cdot |U|^2)$, 故求约简的时间复杂度为 $O(|A|^3 \cdot |U|^2)$;

计算知识粒度 $\text{GK}(A)$ 和 $\text{GK}(A - \{a\})$ 的时间复杂度为 $O(|U|)$ 。

综上可知该算法的最坏时间复杂度为 $O(|A|^3 \cdot |U|^2)$ 。

5 实例分析

为说明算法的有效性, 我们利用文[7]给出的小汽车信息表求最小约简, 见表 1。

表 1

Car	Price	Mileage	Size	Max-Speed
1	High	High	Full	Low
2	Low	*	Full	Low
3	*	*	Compact	High
4	High	*	Full	High
5	*	*	Full	High
6	Low	High	Full	*

这里为书写方便, 不妨用 P, M, S, X 依次代替 Price, Mileage, Size, Max-Speed。

下面对表 1 求最小约简:

步骤 1: 计算知识粒度得 $\text{GK}(A) = 12/36$ 。

(下转第 199 页)

战性问题,特别是当图像中含有面部附属物等干扰信号时,自动识别更加困难。针对这一问题,本文提出基于事例的 AU 识别算法。该方法无需提取面部特征参数,因而减少了对图像成像环境的要求,具有更好的鲁棒性和实用性。以往研究中,已知 AU 的训练集数据作为本算法的事例集。事例图像的 AU 编码作为解决方案。为了获得一个新的面部图像的 AU 编码,算法采用 IGA,通过交互,充分利用人类的比较能力,快速从事例集中检索一个最接近的事例图像,该图像的 AU 编码即成为新图像的 AU,从而实现 AU 的自动识别。为了验证算法的有效性,本文采用简单面部图像和复杂面部图像进行测试,分别获得 77.5% 和 82.8% 平均识别率,以及 82.8% 和 93.1% 平均相似度,比特征脸方法有较大幅度的提高。

表 8 AU 组合的实验结果

图像号码	AU 组合	平均相似度
1	1+2+5+27	50.0%
2	4+17	100.0%
3	1+2+27	66.7%
4	1+4+15+17	25.0%
5	6+12+25	100.0%
6	4+7+9+17	75.0%
7	1+2+26	66.7%
8	15+17	50.0%
9	1+2+5+26	75.0%
10	12+25	0%
11	4+6+7+9+17	40.0%
12	6+12	0%
13	25	100.0%
14	26	0%
15	6+7+12+25	75.0%
16	12	0%

本文提出的方法不仅为 AU 的自动识别提供了新的解决思路,同时将 IGA 应用到了一个新的研究领域。

虽然文中只对部分 AU 和 AU 组合进行了识别,但是文

中的两个实验具有一定的代表性,可以说明本文方法的有效性、可行性。本文的方法可以推广到任何 AU 和 AU 组合的识别。文中只进行了静态图像的实验,下一步工作是把该方法扩展到动态的图像序列中,并与其他方法进行比较实验。另外,在实验中,我们发现事例集图像的多样性非常重要。如果事例集中没有可匹配的事例,那么识别的效果就会降低。如何交互和动态地增加事例集也是我们下一步的研究目标。

致谢 非常感谢 Dr. Jeffrey Cohn 及其助手为本实验提供了 Cohn-Kanade 数据库,并允许作者使用其中的一些图像作为文章中的图例。其网站是:

<http://vasc.ri.cmu.edu/idb/html/face/facial-expression/index.html>.

参考文献

- Goleman D. Emotion Intelligence. Bantam Books, 1995
- Picard R W. Affective Computing. MIT Press, 1997
- Ekman P, Friesen W. Facial Action Coding System: a Technique for the measurement of facial movement. Palo Alto, CA: consulting Psychologists Press, 1978
- Bartlett M S, Hager J C, Ekman P, et al. Measuring Facial Expression by Computer Image Analysis. Psychophysiology, 1999, 36:253~263
- Donato G, Bartlett M S, Hager J C, et al. Classifying Facial Actions. IEEE trans on Analysis and Machine Intelligence, 1999, 21: 974~989
- Cohn J F, Zlochower A J, Lien J, et al. Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. Psychophysiology, 1999, 36: 35~43
- Tian Y L, Kanade T, Cohn J F. Recognizing Action Units for Facial Expression Analysis. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(2): 97~115
- Kapoor A. Automatic Facial Action Analysis. [Master thesis]. MIT, 2002
- Pantic M, Rothkrantz L J M. Facial Action Recognition for Facial Expression Analysis from Static Face Images. IEEE Transactions on Systems, Man, and Cybernetics -Part B, 2004, 34(3): 1449~1461
- Wang Shangfei, Xue jia. Case-based Facial Action Units Recognition Using Interactive Genetic Algorithm. ACII2005, LNCS3784, 2005, 80~87
- Takagi H. Interactive Evolutionary Computation: Fusion of the Capacities of EC Optimization and Human Evaluation. Proceedings of the IEEE, 2001, 89(9): 1275~1296

(上接第 170 页)

步骤 2: 计算核 $\text{Core}(A)$ 。各属性的重要度为: $\text{Sig}_{A-(X)}(X) = \text{Sig}_{A-(S)}(S) = \text{Sig}_{A-(P)}(P) = 9/36$, $\text{Sig}_{A-(M)}(M) = 0$, 于是 A 的核为 $\text{Core}(A) = \{P, S, X\}$ 。 $\text{GK}(\text{Core}(A)) = 12/36$, 故 $\text{Core}(A)$ 为该系统的核心约简, 输出 $\text{Core}(A) = \{P, S, X\}$, 算法终止。

此结果与文[7]利用区分函数得到的约简结果是一致的。不过需要指出的是, 我们给出的算法是尽可能找出不完备信息系统的一个约简, 但一般情况下, 约简不是唯一的, 最小约简也不是唯一的。

结束语 属性约简是粗糙集理论在数据分析上的一个重要应用。一般来讲, 属性约简不惟一。为解决求约简的 NP 问题, 粗糙集研究者们已提出了多种启发式算法。本文针对不完备信息系统, 提出了一种基于知识粒度的属性约简算法, 并分析了该算法的时间复杂度。最后通过例子说明, 使用该算法能找到不完备信息系统的一个约简。

参考文献

- Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341~356

- Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data[M]. Dordrecht: Kluwer Academic Publisher, 1991
- Miao Duoqian, Wang Jue. An information-based algorithm for reduction of knowledge[A]. IEEE ICIPS'97[C], 1997. 1155~1158
- 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
- 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
- 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
- Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences 1998, 112(1-4): 39~49
- Kryszkiewicz M. Rule in incomplete information systems[J]. Information Sciences, 1999, 113(3-4): 271~292
- Liang J Y, Xu Z B. The algorithm on knowledge reduction in incomplete information systems. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 24(1): 95-103
- 何伟, 刘春亚, 赵军, 李华. 不完备信息系统下的属性约简算法[J]. 计算机科学, 2004, 31(2): 117~119
- 黄兵, 周献中, 张蓉蓉. 基于信息量的不完备信息系统属性约简[J]. 系统工程理论与实践, 2005, 25(4): 55~60
- 梁吉业, 李德玉. 信息系统中的不确定性与知识获取[M]. 北京: 科学出版社, 2005. 1~70