

一种基于分辨函数的属性约简算法及其应用

吴茜媛, 郑庆华, 刘广东

(西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 针对个性化网络学习中影响学习策略的关键个性特征发现问题, 基于粗糙集理论, 提出了一种采用分辨函数进行属性约简的算法. 应用分辨函数原理, 逐次比较论域中的对象, 以获取核集和非核集; 利用逻辑运算求取最佳约简, 并对约简集合的属性进行影响因子的排序, 提取出关键个性特征属性. 在西安交通大学英语学习网站平台上的实验结果表明, 所提算法得出的学习者关键个性特征维数仅为原始维度的 $1/4$, 并可自动发现学习者关键个性特征知识, 揭示个性特征和学习策略间的客观关系, 精简 50% 的数据量.

关键词: 网络学习; 个性特征; 学习策略; 分辨函数; 属性约简

中图分类号: TP391 **文献标识码:** A **文章编号:** 0253-987X(2008)12-1455-04

An Attribute Reduction Algorithm to Find Learner's Key Characteristics Based on the Discernible Function

WU Xiyuan, ZHENG Qinghua, LIU Guangdong

(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To find key characteristic attributes, which have deep influence on learning strategies in network learning, an attribute reduction method is proposed based on rough set theory. The method applies the discernible function principle, constructs core set and nocore set to find the reduction through logic calculation, and sorts the attributes in the reduction set by significance factors to obtain the key characteristic attributes. The method is applied to the network English learning platform of Xi'an Jiaotong University and the result shows the dimensions of key personality characteristics are one-quarter of the original ones. The method can find the knowledge of key attributes and reveal the relation between personality characteristic and learning strategy. Based on the results, the original data volume can be decreased to about 50%.

Keywords: network learning; personality characteristic; learning strategy; discernible function; attribute reduction

构建个性化网络学习环境的核心问题之一是制定个性化的学习策略. 教育心理学领域的研究表明, 学习策略的应用, 在很大程度上受学习者个性特征的影响^[1], 而这种特征是指学习者个体内在身心结构和外在行为习惯上表现出来的相对稳定的特点.

目前, 国内外对学习者的个性特征与学习策略之间的依赖关系研究甚多, 但大多数停留在教育心理

学领域, 多采用统计学习的方法, 如相关性、判别式等分析方法^[2-5]. 这些方法只能发现两者间的相关性, 其中哪些个性特征属性是冗余的、可以被去掉的, 还需要领域专家根据显著性水平、支持度和置信度等来判断, 具有一定的主观性. 此外, 相关研究大多数只是简单地分析单个个性特征属性对学习策略的影响, 未考虑多属性组合对学习策略的影响, 更没

有考虑哪些个性特征属性更为重要。

粗糙集作为一种处理不精确、不确定与不完全数据的数学理论^[6],能从数据本身揭示出客观的联系和规则,为分析属性间的依赖关系提供了有效的数学手段。

本文工作主要是:引入粗糙集理论来分析大量学习者的个性特征和学习策略信息,在约简与学习策略无关的个性特征属性的基础上,自动发现学习者关键个性特征知识,进而揭示出个性特征与学习策略间的客观关系。

1 相关概念和定义

定义1 学习者个性特征是学习者个体内在身心结构和外在行为习惯上表现出来的相对稳定的特点,可以表示为 $C = (F_P, s_L, N_L)$,其中 F_P 是刻画人格因素的属性集合, s_L 是刻画学习风格的属性集合, N_L 是刻画学习观念的属性集合。

定义2 学习策略可以表示为 $D = \{l_{gl}, l_{qg}, l_{yx}, l_{yg}, l_{ij}\}$,其中 l_{gl} 代表管理策略属性, l_{qg} 代表情感策略属性, l_{yx} 代表语言形式策略属性, l_{yg} 代表语言功能策略属性, l_{ij} 代表交际策略属性。

对最基本的学习策略的定义和分类,目前在教育学领域依然存在着争论。本文综合了国内外学习策略的划分情况,结合中国学生的实际情况和西安交通大学在外语教学实践中积累的经验认为,将学习策略分为以下4个方面比较恰当:①管理策略,即原认知策略;②情感策略;③认知策略,即学习方法或技巧,它又可以分为语言形式策略和语言功能策略;④交际策略^[5]。

定义3 学习者个性特征信息系统是学习者对象集合,每个学习者对象由属性集合来描述,可以表示为四元组 $S = (U, C \cup D, V, f)$ 。其中: U 为非空有限集,是学习者对象的全体,即论域; C 为非空有限集,是学习者个性特征属性集合; D 为非空有限集,是学习策略属性集合, $C \cap D = \emptyset$, $C \cup D = A$; $V = \bigcup_{a \in A} V_a$, V_a 是 $a \in A$ 的值域; $f: U \times A \rightarrow V$ 是一单射函数,使论域 U 中的任一对象取属性 $a \in A$ 在 V 中的某个惟一值。

此信息系统是分析学习者关键个性特征问题的基础。

定义4 个性特征约简是指,根据个性特征与学习策略间的隐含关系,在去除对学习策略没有影响的个性特征属性后保留下来的个性特征的属性集合。

设 $R \subseteq C$, $X \subseteq U$, 则称 $RX = \{x \in U | [x]_{\text{ind}(R)} \subseteq$

$X\}$ 为 X 的 R -下近似集,其中 $\text{ind}(R)$ 表示 R 上的不可分辨关系,是一种等价关系,通过一个不可分辨关系可以得到信息系统的划分^[6]。 $\text{pos}_R(D) = \bigcup_{X \in U/D} (RX)$ 为 D 的 R 正域,其实质是根据学习者个性特征子集合 R ,学习者对象论域中所有能归入集合 U/D 的元素构成的并集。如果 $\alpha \in R$, $\text{pos}_R(D) = \text{pos}_{R-\{\alpha\}}(D)$, 则 α 为 R 中 D 冗余的;否则,称 α 为 R 中 D 必要的。若在 R 中每个属性 α 都是必要的,则称 R 为 D 独立的。

对于一个给定的学习者个性特征信息系统 S , 个性特征属性集合 C 的约简是 C 的一个非空子集 C' , 它满足 $\text{pos}_{C'}(D) = \text{pos}_C(D)$ 和 C' 为 C 的 D 独立子集,则称 C' 为 C 的 D 相对独立约简,简称约简。 $r_D(C)$ 表示由所有约简构成的集合。

定义5 关键个性特征是指对学习策略的影响,其重要程度高的个性特征属性。若 $\alpha \in C'$, α 的影响因子 $F_{SG}(\alpha, C', D) = (|\text{pos}_{C'}(D)| - |\text{pos}_{C'-\{\alpha\}}(D)|) / |U|$ 。学习策略对个性特征属性(除 α 外)的依赖度为 $1 - F_{SG}(\alpha, C', D)$, 若依赖度小于阈值,则说明 α 对学习策略的影响是关键性的。阈值的设定目前还是一经验值,一般设为 0.97。

定义6 若 $a_i \in C$, $x, y \in U$, 且属于 d_i 的不同等价类,则 $a(x, y) = \{a_1, a_2, \dots, a_k\}$ 表示能区分学习者论域中对象 x, y 的条件属性集合, $\sum a(x, y) = a_1 \vee a_2 \vee \dots \vee a_k$ 。若 $a(x, y) = \emptyset$, 令 $\sum a(x, y) = 1$ 。由此,定义分辨函数为 $\Delta = \prod_{(x, y) \in U \times U} \sum a(x, y)$ 。

定理1 分辨函数 Δ 的最小简化的析取范式对应于 S 的全体约简^[6]。

定义7 个性特征约简前,信息系统 S 的数据量 $E_S = |C| |U|$, 约简后的数据量 $E'_S = |C'| |U|$, 则数据约简效率为 $E = 1 - E'_S / E_S$ 。

2 学习者关键个性特征发现算法

利用粗糙集理论进行约简,国内外研究者已经做了一些工作。Wond 等人已证明了求最小约简是 NP 难问题。若直接使用基于定义4的数据分析方法,在理论上是可行的,但在实际计算中,属性的组合爆炸将造成时间和空间的代价很高,这是不现实的。Skowron 提出了分辨矩阵和分辨函数的概念,用它可以方便地求取属性核和约简^[7-8]。一些研究者基于分辨矩阵也提出了很多种约简方法,但这类方法首先要生成整个论域的分辨矩阵,这势必造成时

间和空间上的浪费。

本文提出了一种方法,它基于分辨函数原理,直接构造核集和非核集,再利用逻辑运算求取最佳约简,最后在最佳约简的基础上,使用定义 5 中的度量 $F_{SG}(a,C',D)$ 对约简集合的属性进行影响因子的排序,提取关键个性特征属性。

核集 ζ_{core} 是指包含惟一能区分论域中任意两个学习者对象的条件属性集合^[6],即 $\zeta_{core} = \{c \in C; \exists \text{ 惟一 } c, \text{ 使得 } \forall c^* \in C, c^* \neq c, \text{ 有 } f(x_i, c^*) = f(x_j, c^*) \text{ 且 } f(x_i, c) \neq f(x_j, c)\}$ 。

非核集 ζ'_{nocore} 是指包含除核集元素外,能区分论域中任意两个学习者对象的其他条件属性的集合,即 $\zeta'_{nocore} = \{a \in C; a \notin \zeta_{core}, \text{ 有 } f(x_i, a) \neq f(x_j, a)\}$ 。由此有 $\zeta_{core} \subseteq C, \zeta'_{nocore} \subseteq C, \zeta_{core} \cap \zeta'_{nocore} = \emptyset$ 。那么,学习者关键个性特征发现算法描述如图 1 所示。

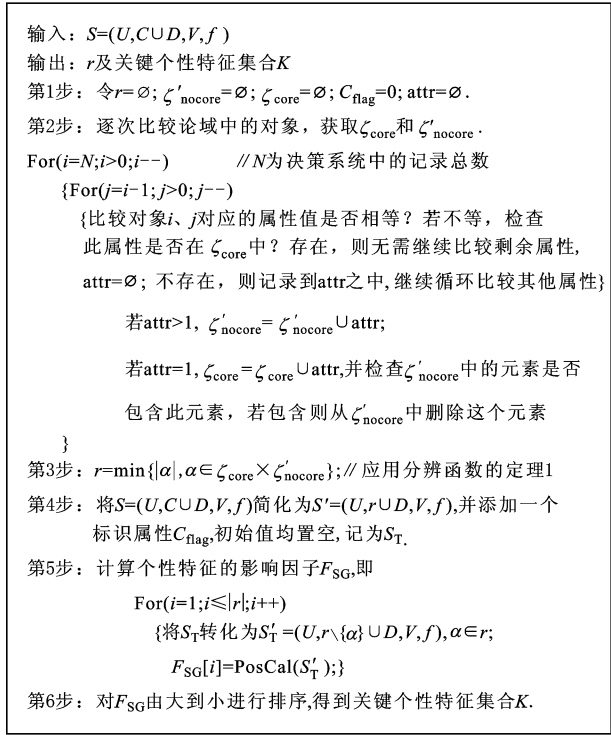


图 1 学习者关键个性特性发现算法

图中,函数 PosCal 的计算可参考文献[9]。由于不需要生成整个论域的分辨矩阵,并且在算法上对包含核属性 ζ'_{nocore} 中的元素提前进行了是否吸收的判断,故与基于分辨矩阵的算法相比,空间复杂度大大降低。在时间复杂度方面,最坏情况下,若每次比较所考虑的属性数都为最大数 $|C|$ ($|C|$ 为个性特征属性个数),则算法的时间复杂度为 $O(N^2 |C|)$ 。

3 实际应用与分析

通过西安交通大学的个性化英语学习网站平台,采集了西安交通大学的理学院、能源与动力工程学院、电气学院、电子与信息工程学院、环境与化学工程学院、软件学院等各专业共约 300 多个学习者的个性特征和学习策略信息,提取有效样本数为 157。

学习者个性特征集合 $C=(F_P, s_L, N_L)$ 包含了 28 维,其中 F_P 有 16 维,是刻画学习者的 Cattell 16FP 人格因素个性特征的; s_L 有 6 维,是刻画学习者的学习风格个性特征的; N_L 有 6 维,是刻画学习者的学习观念个性特征的。

学习策略集合 $D=\{l_{gl}, l_{qg}, l_{yx}, l_{yg}, l_{ij}\}$ 包括 5 个属性,可以将 $S=(U,C \cup D,V,f)$ 分解成 5 个不同的信息系统 $\{S_1, S_2, S_3, S_4, S_5\}$, 其中 $S_i=(U, R_i, V, f), R_i=C \cup \{d_i\}, d_i(1 \leq i \leq 5)$ 代表 $D=\{l_{gl}, l_{qg}, l_{yx}, l_{yg}, l_{ij}\}$ 的第 i 个学习策略属性。对 5 种学习策略分别分析,能更准确地提取出学习者对某种学习策略的偏好。后续实验的结果也分别验证了分析方法的正确性。另外,5 种学习策略是可以单独实施的,分别考量也有利于学习策略在个性化网络学习环境中的设计和实现。

对 $S_i=(U, R_i, V, f)(1 \leq i \leq 5)$,运用第 2 部分的算法进行约简,目标是尽可能去掉多余的个性特征属性,故在多个约简中,选择约简集合的基最小的约简。对有多个最小约简的,只需要取其中的一个即可。可以在产生的规则最少和与 5 种学习策略关系最小中进行选择,得到 5 种学习策略最终的最佳约简见表 1。个性特征对 l_{gl} 的最佳约简的 $F_{SG}(a,C,D)$ 如表 2 所示。

表 1 5 种学习策略最终的个性特征最佳约简	
学习策略	最佳约简的属性集合
l_{gl}	$\{C, E, H, I, M, N, O, Q_2, Q_4, G_1, G_3, G_4, L_M, L_Z, L_L\}$
l_{qg}	$\{B, E, G, H, I, N, O, Q_2, Q_3, Q_4, G_2, G_3, L_G, L_Z, L_C\}$
l_{yx}	$\{E, I, Q_2, G_4, L_W\}$
l_{yg}	$\{A, C, G, H, N, O, Q_2, Q_3, Q_4, G_1, G_3, G_6, L_M, L_W, L_G\}$
l_{ij}	$\{A, C, F, G, I, M, O, Q_4, G_1, G_2, G_3, G_4, G_5, L_G, L_Z, L_C\}$

表2 个性特征对 l_{gl} 的最佳约简的 $F_{SG}(a,C,D)$

个性特征(对 l_{gl})	属性名	$F_{SG}(a,C,D)$
第一组风格(视觉 听觉 实践型)	G_1	0.114 6
世故性	N	0.076 4
敢为性	H	0.070 1
母语依赖观念	L_M	0.063 7
语言学习性质观念	L_Z	0.051 0
敏感性	I	0.038 2
第三组风格(群体 个体型)	G_3	0.038 2
忧虑性	O	0.025 5
独立性	Q_2	0.025 5
紧张性	Q_4	0.025 5
稳定性	C	0.012 7
恃强性	E	0.012 7
学习策略观念	L_C	0.012 7
幻想性	M	0.000 9
第四组风格(分析 总体型)	G_4	0.000 9

从表2可以看出, $G_1、N、H、L_M、L_Z、I、G_3$ 是 $l_{gl} \in D$ 的关键个性特征属性. 对其他的4种学习策略进行同样分析,可以得到 $L_G、I、G_2、H、Q_4、G、Q_2$ 是 $l_{qg} \in D$ 的关键个性特征属性; $E、G_4、L_Z、I、Q_2$ 是 $l_{yx} \in D$ 的关键个性特征属性; $H、O$ 是 $l_{yg} \in D$ 的关键个性特征属性; $I、G_1$ 是 $l_{ji} \in D$ 的关键个性特征属性.

约简前 $S_1、S_2、S_3、S_4、S_5$ 的 E_S 均为4 396,约简后 E'_S 分别为2 355、2 355、785、2 355、2 512. E 分别为46%、46%、82%、46%、43%,可见信息系统数据量的减少是显著的.

4 结束语

本文提出的基于粗糙集理论的学习者关键个性特征发现方法,为研究学习者个性特征与学习策略之间的关系提供了客观、量化的数据推理方法,它不需要其他先验知识,如统计概率分布或模糊集合论的隶属度或隶属函数等^[10].

但是,本文算法对最小约简的完备性还有待进一步探讨. 另外,如何应用所获取的个性特征约简和影响因子,进一步建立优化的学习者模型、构建个性化网络学习环境等,需要进行更深入的研究.

参考文献:

[1] 刘电芝,黄希庭. 学习策略研究概述[J]. 教育研究, 2002(2): 81.
LIU Dianzhi, HUANG Xiting. The introduction of learning research [J]. Education Research, 2002(2): 81.

[2] 王振宏,刘萍. 动机因素、学习策略、智力水平对学生学业成就的影响[J]. 心理学报, 2000,32(1): 65-69.
WANG Zhenhong, LIU Ping. The influence of motivational factors, learning strategy, and the level of intelligence on the academic achievement of students [J]. Journal of Chinese Psychology Acta Psychologica Sinica, 2000, 32(1): 65-69.

[3] VERMETTEN J, LODEWIJKS G, VERMUNT D. The role of personality traits and goal orientations in strategy use[J]. Contemporary Educational Psychology, 2001 (26): 150-152.

[4] BUSATO V V, PRINS F J, ELSHOUT J J, et al. The relation between learning styles, the big five personality traits and achievement motivation in higher education [J]. Personality and Individual Differences, 1999,26: 129-140.

[5] 刘润清,戴曼纯. 中国高校外语教学改革现状与发展策略研究 [M]. 北京: 外语教学与研究出版社,2003.

[6] 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, 2001.

[7] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems [M]// Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht, Netherlands: Kluwer Academic Publishers, 1992: 331-362.

[8] 王珏, 王任, 苗夺谦, 等. 基于 Rough Set 理论的“数据浓缩”[J]. 计算机学报, 1998(5): 393-400.
WANG Jue, WANG Ren, MIAO Duoqian, et al. Data enriching approach based on rough set theory[J]. Chinese Journal of Computers, 1998(5): 393-400.

[9] 吴茜媛. 网络学习中学习者个性特征和学习策略依赖关系研究 [D]. 西安交通大学电子与信息工程学院, 2005.

[10] PAWLAK Z, GRZYMALA-BUSSE J, SLOWINSKI R, et al. Rough sets [J]. Communications of the ACM, 1995,38 (11): 89-95.

(编辑 苗凌)