

DOI:10.13232/j.cnki.jnju.2019.04.001

属性的变化对于流图的影响

姚 宁^{1,2}, 苗夺谦^{1,2*}, 张远健^{1,2}, 康向平^{1,2}

(1. 同济大学计算机科学与技术系, 上海, 201804;

2. 同济大学嵌入式系统和服务计算教育部重点实验室, 上海, 201804)

摘 要: 人类的认知中具有粒化特性, 并且同一现象在不同粒度上具有不同的解释. 流图为知识的一种表示形式, 素有直观性、计算便捷性和并行处理等特征. 以属性-值形式的信息系统作为研究对象, 针对新属性的添加而诱导的粒度变化, 研究流图在不同粒度上的具体演变. 流图在新粒度上的有效性取决于所涉及的等价类的变化和 Markov 性质的成立. 具体的, 若新粒度上仅有部分等价类中的成员保持 Markov 性质成立, 则粒度变化可将图形结构由一个粒度上的流图转化为新粒度上的用于构成完整流图的基本构件; 若 Markov 性质在新粒度上不成立, 则流图可被转化为新粒度上的与流图无关的结构; 若新粒度上等价类中的每个成员皆满足 Markov 性质, 则流图在新粒度上保持不变. 流感病人信息系统在不同粒度上的具体分析进一步验证了理论结果. 这些结论有助于理解和刻画知识与粒度之间的关系, 为模拟人类学习和思维奠定基础.

关键词: 流 图, Markov 性质, 等价类, 粒 度, 粗糙集

中图分类号: TP18

文献标识码: A

The impact of changing attributes on flow graph

Yao Ning^{1,2}, Miao Duoqian^{1,2*}, Zhang Yuanjian^{1,2}, Kang Xiangping^{1,2}

(1. Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China;

2. Key Laboratory of Embedded System & Service Computing, Ministry of Education of China,
Tongji University, Shanghai, 201804, China)

Abstract: Granulation is an inherent property of human cognition and the same phenomenon has different interpretations at different granularities. Flow graph is treated as a form of knowledge representation, and is known for its intuitive formation, straightforward computation and parallel processing. Taking the attribute-value information system as the research object, this paper studies the specific changes of the flow graph at different granularities which are induced by adding the new attribute(s). The validity of the flow graph at a new granularity depends on the change of the equivalence classes involved and the establishment of the Markov property. Specifically, if only parts of elements of the equivalence class at the new granularity maintain the Markov property, the change in granularity will then cause the graphical structure to be transformed from a flow graph at one granularity to a basic component for a complete flow graph at this new granularity. If the Markov property does not hold at the new granularity, the flow graph will be transformed into a

基金项目: 国家重点研发计划(213), 国家自然科学基金(61673301, 61573255, 61573259, 61673299, 61603278), 公安部重大专项(20170004)

收稿日期: 2019-05-17

* 通讯联系人, E-mail: dqmiao@tongji.edu.cn

structure that is unrelated to flow graph at this new granularity. If every element of the equivalence class at the new granularity satisfies the Markov property, the flow graph at one granularity will then remain unchanged at this new granularity. The illustrations of an information system on patients suffering from flu at different granularities further validate the proposed theoretical results. These conclusions can help to understand and characterize the relationship between knowledge and granularity, and lay the foundation for simulating human learning and thinking.

Key words: flow graph, Markov property, equivalence class, granularity, rough set

数据作为最重要的资源之一,在日常推理、机器学习、人工智能等研究中发挥着关键性作用.将数据作为研究的出发点,不需要关于数据的任何预备的或附加的信息,直接关注数据本身所能表示和传达的信息或知识,正是粗糙集理论在数据分析中所具有的优势和独特之处.

粗糙集理论由 Pawlak^[1]依据 Frege^[2](德国哲学家、逻辑学家、数学家、现代逻辑之父)的“模糊性关联于边界区域,即存在不能唯一被归类于一个集合或者其补集的对象”的观点的具体实现而提出,开始于形式化知识表示系统或信息系统或属性-值数据表格.该理论认为数据自身的不可区分性导致数据具有粒状结构,进而可被用于挖掘数据中蕴涵的信息^[3].图表示的核心也是将数据或事实归类于结构,通过这些结构,所找寻的信息可与任务中涉及的变量关联起来,沿着图中展示的路径,人类的许多推理模式可得到阐述. Pawlak^[4]还提出了粗糙集理论中的流图表示,将条件粒(条件属性诱导的数据的粒状结构)和决策粒(决策属性诱导的数据的粒状结构)之间的关系描述为流网络(Flow networks),流图中的信息流由确定因子和覆盖因子(二者皆为条件概率的粗糙集术语描述的对应用)决定,并受全概率公式和 Bayes 定理支配.这类流图关注于网络中的流分布建模,不同于 Ford and Fulkerson^[5]提出的流图(聚焦于图中的最优流分析).流图的这种形式上的直观性、计算上的简单便捷性和并行处理特性,使得基于流图的推理相比于直接从数据做推理更直观、更具可解释性.流图的研究成果主要集中于流图的构建(比如基于等价类^[6]或者基于决策树^[7])和流图中系数的确定(比如利

用因式分解逐个消除变量^[8]或者基于矩阵的表示形式^[9]以多项式时间计算流图中的系数),近年来在应用方面粗糙集流图在多目标优化问题近似求解中展示了重要优势^[10].

Yao and Miao^[6]将流图解释为一类特殊的概率图模型,即具有 Markov 性质的有向非循环图(Directed acyclic graphs),从而基于属性诱导的等价类直接由数据构建流图.该图形结构描述变量间的一种层次架构,不同的属性值对应不同层,每层中节点间满足 Markov 性质,整个架构中两终端节点之间的条件概率满足全概率公式.姚宁等^[11]还探讨了因果信息对于属性变化而产生的迁移问题.本文是这些工作的延续,研究信息系统中属性的变化对于已抽取的流图的影响,或者流图在不同信息系统中的可迁移性.这里属性的变化侧重于属性个数的增加或删减,关于属性值层面的粒度变化的最新研究成果可参考文献[12].考虑到属性的变化与系统所能刻画的信息的粒度,即信息粒中所包含的成員的数量直接相关,文中给出的结论可窥见粒度变化与流图变化之间的关联,为理解和刻画人类认知中的粒化属性提供新的视角.

1 基本概念与相关工作

本节主要回顾粗糙集理论中与流图有关的基本概念和已有成果.

定义 1 流图^[7] 流图为一有限的有向非循环图,记作 $G=(N,B,\varphi)$,其中 N 表示一节点集, $B\subseteq N\times N$ 表示一有向分支集, $\varphi:B\rightarrow R^+$ 表示一个值域为非负实数的流函数.

(1) 对于任意的 $w,y\in N$,若 $(w,y)\in B$,

则称 w 为 y 的一个输入, 对应的, y 为 w 的一个输出; 称 $\varphi(w, y)$ 为从 w 到 y 的一个流函数 (也有将 φ 称作 a throughflow) 并且假设 $\varphi(w, y) \neq 0$; 所有 w 的输入集记作 $I(w)$, 所有 w 的输出集记作 $O(w)$. 流图 G 的输入和输出, 分别记作 $I(G)$ 和 $O(G)$, 定义如下:

$$I(G) = \{w \in N: I(w) = \emptyset\}$$

$$O(G) = \{w \in N: O(w) = \emptyset\}$$

称 G 的输入和输出为 G 的外部节点, G 的其他节点为 G 的内部节点. 流图 G 的流函数 (throughflow), 记作 $\varphi(G)$, $\varphi_{\text{inflow}}(G)$ 表示 G 的输入流函数 (简称为输入流), $\varphi_{\text{outflow}}(G)$ 表示 G 的输出流函数 (简称为输出流), 定义为:

$$\varphi_{\text{outflow}}(G) = \sum_{w \in I(G)} \sum_{y \in O(w)} \varphi(y, w) =$$

$$\varphi(G) = \sum_{w \in I(G)} \sum_{y \in O(w)} \varphi(w, y) = \varphi_{\text{inflow}}(G)$$

相应的, 节点 w 的流函数记作 $\varphi(w)$, 节点 w 的输入流和输出流分别记作 $\varphi_{\text{inflow}}(w)$ 和 $\varphi_{\text{outflow}}(w)$, 具体定义如下:

$$\varphi_{\text{outflow}}(w) = \sum_{y \in O(w)} \varphi(w, y)$$

$$\varphi_{\text{inflow}}(w) = \sum_{y \in I(w)} \varphi(y, w)$$

当 w 为流图 G 的内部节点时, 有:

$$\varphi_{\text{outflow}}(w) = \varphi(w) = \varphi_{\text{inflow}}(w)$$

令 $\sigma: B \rightarrow [0, 1]$, 对于每个 $(w, y) \in B$, 定义 $\sigma(w, y) = \frac{\varphi(w, y)}{\varphi(G)}$, 称 $\sigma(w, y)$ 为 (w, y) 的强度, 则得到一正规化 (normalized) 流图, 记作 $G = (N, B, \sigma)$.

(2) G 中由 w 到 y ($w \neq y$) 的一条 (有向) 路 (a (directed) path) 定义为一节点序列 w_1, \dots, w_n , 其中 $w_1 = w$, $w_n = y$. 对于每个 $i (1 \leq i \leq n-1)$, $(w_i, w_{i+1}) \in B$, 采用符号术语记作 $[w \cdots y]$. G 中从 w 到 y 的所有路构成的集合记作 $\langle w, y \rangle$, 称为 G 中从 w 到 y 的联结 (a connection). 对于 G 中每个分支 (w, y) , 指定两个数学量: 确定因子 (certainty factor) 和覆盖因子 (coverage factor), 分别记作 $\text{cer}(w, y)$ 和

$\text{cov}(w, y)$, 具体定义式为:

$$\text{cer}(w, y) = \frac{\sigma(w, y)}{\sigma(w)}$$

$$\text{cov}(w, y) = \frac{\sigma(w, y)}{\sigma(y)}$$

并且, $\sigma(w) \neq 0, \sigma(y) \neq 0$. 相应的, 路 $[w_1, \dots, w_n]$ 的确定因子和覆盖因子定义如下:

$$\text{cer}[w_1, \dots, w_n] = \prod_{i=1}^{n-1} \text{cer}(w_i, w_{i+1})$$

$$\text{cov}[w_1, \dots, w_n] = \prod_{i=1}^{n-1} \text{cov}(w_i, w_{i+1})$$

联结 $\langle w, y \rangle$ 的确定因子和覆盖因子有如下定义:

$$\text{cer}\langle w, y \rangle = \sum_{[w, \dots, y] \in \langle w, y \rangle} \text{cer}[w, \dots, y]$$

$$\text{cov}\langle w, y \rangle = \sum_{[w, \dots, y] \in \langle w, y \rangle} \text{cov}[w, \dots, y]$$

定义 2 等价类^[13] 设 $U \neq \emptyset$ 为一有限的对象集合, 称作全域, R 为 U 上由对象的属性诱导的一个等价关系. 对于 U 上任意元素 $x \in U$, x 关于等价关系 R 的等价类, 记作 $[x]_R = \{y \in U: xRy\}$, 称为 R 中的一个概念或者类别, 有时也称作一个粒子. U/R 表示 R 的所有等价类组成的集族, 称作与 R 有关的知识 (简称知识 R), 其构成了 U 的一个划分.

定义 3 确定因子^[4, 6] 设 (U, A) 为一信息系统, 其中有限非空集合 U 为对象集, 称作全域; 有限非空集合 A 称作属性集. 若将属性集 A 区分为条件属性集 C 和决策属性集 D , 则称 (U, C, D) 为一决策表, 其中 $C = \{C_1, \dots, C_n\}$, $D = \{D_1, \dots, D_m\}$, n 和 m 为非负整数. 对于任意 $x \in U$, 由 x 诱导的决策规则定义为一序列 $[x]_{C_1}, \dots, [x]_{C_n}, [x]_{D_1}, \dots, [x]_{D_m}$, 记作 $[x]_{C_1}, \dots, [x]_{C_n} \rightarrow [x]_{D_1}, \dots, [x]_{D_m}$ (简记 $C \rightarrow_x D$). 决策规则确定因子, 记作 $\text{cer}_x(C, D)$, 具体定义式为:

$$\text{cer}_x(C, D) = \frac{|[x]_C \cap [x]_D|}{|[x]_C|}$$

其中, $|X|$ 表示集合 X 的基数, 即集合 X 中元素

的个数. 决策规则的确定因子可从统计学上(频率)解释为一条条件概率, 即已知 y 属于 $[x]_C$, 则 y 属于 $[x]_D$ 的条件概率, 用符号术语标记为 $P_x(D|C)$. 这里为书写方便, 可省略下标 x , 直接写作 $C \rightarrow D$, $cer(C, D)$ 和 $P(D|C)$. 决策规则的覆盖因子, 记作 $cov(C, D)$, 具体定义为:

$$cov(C, D) = \frac{|[x]_C \cap [x]_D|}{|[x]_D|} = P(C|D)$$

可解释为已知决策属性 D , 则 x 属于条件类(由条件属性诱导的等价类) $[x]_C$ 的条件概率. U 中对象 x 满足条件属性 C 的边缘概率(无条件概率), 记作 $P(C)$, 定义为:

$$P(C) = \frac{|[x]_C|}{|U|}$$

决策规则的强度记作 $\sigma(C, D)$, 定义为:

$$\sigma(C, D) = \frac{|[x]_C \cap [x]_D|}{|U|}$$

若 $cer(C, D) = 1$ (或者 $P(D|C) = 1$), 则称 $C \rightarrow D$ 为系统中的确定决策规则; 若 $0 < cer(C, D) < 1$ (或者 $0 < P(D|C) < 1$), 则称 $C \rightarrow D$ 为系统中的不确定决策规则.

注 释 定义 1 中基于流图的确定因子和覆盖因子的概念与定义 3 中基于数据表格的相应概念等价, 并且定义 1 中的 $\varphi(w, y)$ 指全域 U 中满足分支 (w, y) 的对象的个数.

下面给出基于等价类直接从数据中抽取流图的两个主要结论, 即引理 1 和引理 2, 被用于粒度变化中识别流图存在的重要依据.

引理 1^[6] 设 (U, A) 为一信息系统, U 为一非空有限对象集, 称作全域, A 为一有序属性集, 即 $A = \{A_1, \dots, A_n\}$, a_1, \dots, a_n 为对应属性变量 A_1, \dots, A_n 的属性值. 设 $[x]_{a_i}$ 表示由属性 A_i 诱导的包含元素 $x \in U$ 的等价类, U/A_i 表示由属性 A_i 诱导的所有等价类组成的集族, 称作 U 的一个划分. 设符号 $|\cdot|$ 表示集合的基数.

则对于 $P(A_1, \dots, A_{n-1}) > 0$, 若要使得:

$$P(A_n | A_1, \dots, A_{n-1}) = P(A_n | A_{n-1})$$

当且仅当对于属性 A_{n-1} 的每个等价类 $[x]_{a_{n-1}}$ 、由属性组合 $\{A_1, \dots, A_{n-2}, A_{n-1}\}$ 诱导的每个等价类 $[x]_{a_1, \dots, a_{n-2}, a_{n-1}}$ 、由属性组合 $\{A_{n-1}, A_n\}$ 诱导的每个等价类 $[x]_{a_{n-1}, a_n}$ 以及由属性组合 $\{A_1, \dots, A_{n-2}, A_{n-1}, A_n\}$ 诱导的每个等价类 $[x]_{a_1, \dots, a_{n-2}, a_{n-1}, a_n}$, 存在某个系数 $\gamma \geq 1$ 使得这些等价类满足下列等式:

$$\begin{aligned} |[x]_{a_{n-1}}| &= \gamma |[x]_{a_1, \dots, a_{n-2}, a_{n-1}}| \\ |[x]_{a_{n-1}, a_n}| &= \gamma |[x]_{a_1, \dots, a_{n-2}, a_{n-1}, a_n}| \end{aligned}$$

引理 2^[6] 设 (U, A) 为一信息系统, U 为一非空有限对象集, 称作全域, A 为属性集. 设 $x \in U$, 并且 $W, Y, Z \in A$, 属性 W, Y, Z 的属性值分别为 $w = \{w_i: 1 \leq i \leq n\}$, $y = \{y_j: 1 \leq j \leq m\}$, $z = \{z_k: 1 \leq k \leq q\}$ (i, j, k, m, n, q 为正整数). 设 $[x]_{w_i}$ 表示属性 W 取值为 w_i 时诱导的等价类, $[x]_{y_j}$ 表示属性 Y 取值为 y_j 时诱导的等价类, $[x]_{z_k}$ 表示属性 Z 取值为 z_k 时诱导的等价类, U/W 表示由属性 W 诱导的全域 U 的划分, 即:

$$U/W = \{[x]_{w_1}, \dots, [x]_{w_i}, \dots, [x]_{w_n}\}$$

U/Y 表示由属性 Y 诱导的全域 U 的划分, 即:

$$U/Y = \{[x]_{y_1}, \dots, [x]_{y_j}, \dots, [x]_{y_m}\}$$

U/Z 表示由属性 Z 诱导的全域 U 的划分, 即:

$$U/Z = \{[x]_{z_1}, \dots, [x]_{z_k}, \dots, [x]_{z_q}\}$$

(1) 若对于每个等价类 $[x]_{z_i}, [x]_{z_i, y}, [x]_{z_i, w}$ 和 $[x]_{z_i, y, w}$, 存在某个系数 $\gamma \geq 1$ 使得:

$$\begin{aligned} |[x]_{z_i}| &= \gamma |[x]_{z_i, w}| \\ |[x]_{z_i, y}| &= \gamma |[x]_{z_i, y, w}| \end{aligned}$$

则下面等式成立:

$$P(y|w, z) = P(y|z)$$

$$P(y|w) = \sum_z P(y|z)P(z|w)$$

同时属性 W, Z 和 Y 之间流图, 即 $W \rightarrow Z \rightarrow Y$ 存在.

(2)若对于每个等价类 $[x]_z, [x]_{z, y_j}, [x]_{z, w_i}$ 和 $[x]_{z, y_j, w_i}$, 存在某个系数 $\gamma \geq 1$ 使得:

$$|[x]_z| = \gamma |[x]_{z, w_i}|$$

$$|[x]_{z, y_j}| = \gamma |[x]_{z, y_j, w_i}|$$

则下列等式成立:

$$P(y|w_i) = \sum_z P(y|z)P(z|w_i)$$

同时属性 w_i, z 和 y_j 之间流图, 即 $w_i \rightarrow z \rightarrow y_j$ 存在.

注 释 流图中变量之间的缺失箭头表示统计独立性的存在, 或为条件独立或为相互独立; 变量之间的箭头(有向连接)表示变量之间的概率依赖. 流图存在的要求: 对于由具体属性值诱导的等价类, 比如 $[x]_{w_i, y_j}$, 该等价类中每个成员皆满足关于有序三元组 (w_i, z, y_j) 的 Markov 性质, 即等价类 $[x]_{w_i, y_j}$ 中的每个成员皆通过有向路 (w_i, z, y_j) 被输出. 换言之, 流图 $w_i \rightarrow z \rightarrow y_j$ 中输入流和输出流相等, 二者皆为 $[x]_{w_i, y_j}$ 中的所有成员; 若由所有属性值诱导的等价类 $[x]_{w, y}$ 中每个成员皆满足关于有序三元组 (w, z, y) 的 Markov 性质, 即每个成员皆通过有向路 (w, z, y) 被输出, 则可得属性 W, Z 和 Y 之间的流图 $W \rightarrow Z \rightarrow Y$. 这里所出现的中间变量 z 可为属性 Z 对应的全部属性值, 或者为属性 Z 对应的部分属性值, 同时属性 Z 可为单个属性也可为多个属性的组合.

2 新属性的添加对于流图的影响

给定信息系统中的对象, 当有新属性被观测到或被发现时, 系统中由属性诱导的对象集的划分和描述将发生变化, 基于引理 2 给出的原始系统中流图表示将如何变化?

假 设 设变量 W, Y, Z 为一信息系统 (U, A) 中任意三个属性, w, y, z 分别为三个属性对应的属性值, 且属性 W, Y, Z 之间的流图 $W \rightarrow Z \rightarrow Y$ 存在. 设 P_{old} 为系统 (U, A) 中的观测分布. 当向该系统中添加一新属性 S , 原始系统 (U, A) 变为系统 $(U, A \cup S)$, 新系统中的观测分布记为 P .

基于该假设前提, 下面给出新属性的添加对于原始系统中流图的影响的相应结论.

定理 1 在新系统 $(U, A \cup S)$ 中, 若对于每个等价类 $[x]_z, [x]_{z, y}, [x]_{w, z, s}, [x]_{w, z, s, y}, [x]_w, [x]_{z, w}$ 和 $[x]_{w, s}$, 存在系数 $\gamma_1, \gamma_2 \geq 1$ 使得:

$$|[x]_z| = \gamma_1 |[x]_{w, z, s}|$$

$$|[x]_{z, y}| = \gamma_1 |[x]_{w, z, s, y}|$$

$$|[x]_w| = \gamma_2 |[x]_{w, s}|$$

$$|[x]_{z, w}| = \gamma_2 |[x]_{z, w, s}|$$

则有:

$$P(y|w, z, s) = P(y|z)$$

$$P(z|w, s) = P(z|w)$$

此时, 新属性 S 仅作用于 W , 同时原系统中的流图 $W \rightarrow Z \rightarrow Y$ 在新系统中仍成立, 并且新系统中还存在新的流图 $S \rightarrow W \rightarrow Z \rightarrow Y$.

定理 2 在新系统 $(U, A \cup S)$ 中, 若对于每个等价类 $[x]_z, [x]_{z, y}, [x]_{w, z, s}, [x]_{w, z, s, y}, [x]_w, [x]_{z, w}$ 和 $[x]_{w, s}$, 仅存在系数 $\gamma_1, \gamma_2 \geq 1$ 使得:

$$|[x]_z| = \gamma_1 |[x]_{w, z, s}|$$

$$|[x]_{z, y}| = \gamma_1 |[x]_{w, z, s, y}|$$

$$|[x]_z| = \gamma_2 |[x]_{w, z}|$$

$$|[x]_{z, y}| = \gamma_2 |[x]_{w, z, y}|$$

不存在系数 $\gamma_3 \geq 1$ 使得:

$$|[x]_w| = \gamma_3 |[x]_{w, s}|$$

$$\left| [x]_{z,w} \right| = \gamma_3 \left| [x]_{w,z,s} \right|$$

则有:

$$P(y|w,z,s) = P(y|z) = P(y|w,z)$$

$$P(z|w,s) \neq P(z|w)$$

此时,新属性 S 仅作用于 Z 或者 $\{W, Z\}$.

若对于每个等价类 $[x]_z, [x]_{z,y}, [x]_{w,z,s}, [x]_{w,z,s,y}, [x]_{w,z}$ 和 $[x]_{w,z,y}$, 不存在系数 $\gamma_1 \geq 1$ 使得:

$$\left| [x]_z \right| = \gamma_1 \left| [x]_{w,z,s} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_1 \left| [x]_{w,z,s,y} \right|$$

但仅存在系数 $\gamma_2 \geq 1$ 使得:

$$\left| [x]_z \right| = \gamma_2 \left| [x]_{w,z} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_2 \left| [x]_{w,z,y} \right|$$

则有:

$$P(y|w,z,s) \neq P(y|z)$$

但是

$$P(y|w,z) = P(y|z)$$

此时,新属性 S 作用于 Y 或者 $\{Z, Y\}$ 或者 $\{W, Z, Y\}$. 这两种情形中原系统中的流图 $W \rightarrow Z \rightarrow Y$ 在新系统中仍成立,即新系统和原始系统具有相同的流图 $W \rightarrow Z \rightarrow Y$.

定理 3 在新系统 $(U, A \cup S)$ 中,若对于每个等价类 $[x]_z, [x]_{z,y}, [x]_{w,z,s}, [x]_{w,z,s,y}, [x]_{z,w,y}$ 和 $[x]_{z,w}$, 不存在系数 $\gamma_1 \geq 1$ 使得:

$$\left| [x]_z \right| = \gamma_1 \left| [x]_{w,z} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_1 \left| [x]_{w,z,y} \right|$$

则有:

$$P(y|w,z) \neq P(y|z)$$

此时,原系统中的流图 $W \rightarrow Z \rightarrow Y$ 在新系统中将不再成立.

定理 4 在新系统 $(U, A \cup S)$ 中,若对于每个等价类 $[x]_z, [x]_y, [x]_{s,y}, [x]_{z,y}, [x]_{w,z}, [x]_{w,z,y}, [x]_{w,z,s}$ 和 $[x]_{w,z,s,y}$, 存在系数 $\gamma_1, \gamma_2, \gamma_3 \geq 1$

使得:

$$\left| [x]_y \right| = \gamma_1 \left| [x]_{w,z,y} \right|$$

$$\left| [x]_{s,y} \right| = \gamma_1 \left| [x]_{w,z,s,y} \right|$$

$$\left| [x]_z \right| = \gamma_2 \left| [x]_{w,z} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_2 \left| [x]_{w,z,y} \right|$$

$$\left| [x]_z \right| = \gamma_3 \left| [x]_{w,z,s} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_3 \left| [x]_{w,z,s,y} \right|$$

则有:

$$P(s|w,z,y) = P(s|y)$$

$$P(y|w,z) = P(y|z) = P(y|w,z,s)$$

此时,仅属性 Y 作用于 S ,同时原系统中的流图 $W \rightarrow Z \rightarrow Y$ 在新系统中仍成立,并且新系统中还存在新的流图 $W \rightarrow Z \rightarrow Y \rightarrow S$.

定理 5 在新系统 $(U, A \cup S)$ 中,若对于每个等价类 $[x]_z, [x]_{w,z}, [x]_{z,s}, [x]_{z,y}, [x]_{w,z,y}, [x]_{w,z,s}$ 和 $[x]_{w,z,s,y}$, 存在系数 $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \geq 1$ 使得:

$$\left| [x]_z \right| = \gamma_1 \left| [x]_{w,z,y} \right|$$

$$\left| [x]_{z,s} \right| = \gamma_1 \left| [x]_{w,z,s,y} \right|$$

$$\left| [x]_z \right| = \gamma_2 \left| [x]_{w,z} \right|$$

$$\left| [x]_{z,s} \right| = \gamma_2 \left| [x]_{w,z,s} \right|$$

$$\left| [x]_z \right| = \gamma_3 \left| [x]_{w,z} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_3 \left| [x]_{w,z,y} \right|$$

$$\left| [x]_z \right| = \gamma_4 \left| [x]_{w,z,s} \right|$$

$$\left| [x]_{z,y} \right| = \gamma_4 \left| [x]_{w,z,s,y} \right|$$

则有:

$$P(s|y,w,z) = P(s|z) = P(s|w,z)$$

$$P(y|w,z) = P(y|z) = P(y|w,z,s)$$

此时,仅属性 Z 作用于 S ,同时原系统中的流图 $W \rightarrow Z \rightarrow Y$ 在新系统中仍成立,并且新系统中还存在流图 $W \rightarrow Z \rightarrow S$.

定理 6 在新系统 $(U, A \cup S)$ 中, 若对于每个等价类 $[x]_w, [x]_z, [x]_y, [x]_s, [x]_{s,y}, [x]_{s,z}, [x]_{z,y}, [x]_{w,s}, [x]_{w,z}, [x]_{w,y}, [x]_{w,z,y}, [x]_{w,z,s}, [x]_{z,y,s}$ 和 $[x]_{w,z,s,y}$, 首先存在系数 $\gamma_1, \gamma_2 \geq 1$ 使得:

$$\begin{aligned} | [x]_z | &= \gamma_1 | [x]_{w,z} | \\ | [x]_{z,y} | &= \gamma_1 | [x]_{w,z,y} | \\ | [x]_z | &= \gamma_2 | [x]_{w,z,s} | \\ | [x]_{z,y} | &= \gamma_2 | [x]_{w,z,s,y} | \end{aligned}$$

即有:

$$P(y|w,z) = P(y|z) = P(y|w,z,s)$$

成立. 同时,

(1) 若还存在系数 $\gamma_3 \geq 1$,

$$\begin{aligned} | [x]_w | &= \gamma_3 | [x]_{w,z,y} | \\ | [x]_{w,s} | &= \gamma_3 | [x]_{w,z,s,y} | \end{aligned}$$

即有:

$$P(s|w,z,y) = P(s|w)$$

此时, 仅属性 W 作用于 S .

(2) 若还存在系数 $\gamma_4 \geq 1$, 使得:

$$\begin{aligned} | [x]_{w,z} | &= \gamma_4 | [x]_{w,z,y} | \\ | [x]_{w,z,s} | &= \gamma_4 | [x]_{w,z,s,y} | \end{aligned}$$

即有:

$$P(s|w,z,y) = P(s|w,z)$$

此时, 仅 $\{W, Z\}$ 作用于 S .

(3) 若还存在系数 $\gamma_5 \geq 1$, 使得:

$$\begin{aligned} | [x]_{z,y} | &= \gamma_5 | [x]_{w,z,y} | \\ | [x]_{z,y,s} | &= \gamma_5 | [x]_{w,z,s,y} | \end{aligned}$$

即有:

$$P(s|w,z,y) = P(s|z,y)$$

此时, 仅 $\{Z, Y\}$ 作用于 S .

(4) 若还存在系数 $\gamma_6 \geq 1$, 使得:

$$\begin{aligned} | [x]_{w,y} | &= \gamma_6 | [x]_{w,z,y} | \\ | [x]_{w,y,s} | &= \gamma_6 | [x]_{w,z,s,y} | \end{aligned}$$

即有:

$$P(s|w,z,y) = P(s|w,y)$$

此时, 仅 $\{W, Y\}$ 作用于 S .

(5) 若存在系数 $\gamma_7, \gamma_8, \gamma_9, \gamma_{10}, \gamma_{11}, \gamma_{12}, \gamma_{13} \geq 1$ 使得:

$$\begin{aligned} |U| &= \gamma_7 | [x]_w |, | [x]_s | = \gamma_7 | [x]_{w,s} | \\ |U| &= \gamma_8 | [x]_z |, | [x]_s | = \gamma_8 | [x]_{s,z} | \\ |U| &= \gamma_9 | [x]_y |, | [x]_s | = \gamma_9 | [x]_{s,y} | \\ |U| &= \gamma_{10} | [x]_{w,y} |, | [x]_s | = \gamma_{10} | [x]_{w,y,s} | \\ |U| &= \gamma_{11} | [x]_{w,z} |, | [x]_s | = \gamma_{11} | [x]_{w,z,s} | \\ |U| &= \gamma_{12} | [x]_{z,y} |, | [x]_s | = \gamma_{12} | [x]_{z,y,s} | \\ |U| &= \gamma_{13} | [x]_{w,z,y} |, | [x]_s | = \gamma_{13} | [x]_{w,z,y,s} | \end{aligned}$$

即有:

$$\begin{aligned} P(s|w,z,y) &= P(s|w,z) = P(s|z,y) = \\ &= P(s|w,y) = P(s|w) = P(s|y) = \\ &= P(s|z) = P(s) \end{aligned}$$

此时, 新属性 S 与 W, Y, Z 皆相互独立.

(6) 或者若不存在系数使得 $P(s|w,z,y)$ 与 $P(s|w,y), P(s|z,y), P(s|w,z), P(s|w), P(s|z), P(s|y)$ 中的任何一个相等, 此时有 $\{W, Z, Y\}$ 作用于 S .

此六种情形皆满足: 原系统中流图 $W \rightarrow Z \rightarrow Y$ 在新系统中仍成立, 即新系统和原始系统具有相同的流图 $W \rightarrow Z \rightarrow Y$.

注 释 定理 1 至定理 6 的证明可直接依据引理 1 和引理 2 证得, 此处省略具体证明过程. 新属性 S 可看作单个属性或多个属性的组合; 若有多个新属性同时被观测到, 则可依据所需一次处理一个新属性或任意个新属性的组合, 加之并行化处理. 定理中所涉及的系数

$\gamma_i (i=1, \dots, 13)$ 可取相同的值也可为不同的数值, 只要确保数值大于或等于 1 即可.

定理 1 至定理 6 详细分析了原始系统中流图在新系统中的有效性以及新属性对于流图自身的形状所产生的具体变化. 由引理 1 和引理 2 可知流图的识别涉及等价类和 Markov 性质. 新属性的增加会引起属性关于全域的划分发生变化, 进而等价类中所含元素也会受到影响, 使得原始系统中的流图在新系统中可能变为流图中的基本构建模块. 因此流图在新系统中的变化需要同时考虑等价类和 Markov 性质, 比如系统 (U, A) 中的流图 $w_i \rightarrow z_k \rightarrow y_j$ 在新系统 $(U, A \cup S)$ 中仍成立, 需要满足在新系统中:

$$P(y_j | w_i, z_k) = P(y_j | z_k)$$

并且

$$[x]_{w_i, y_j} = [x]_{w_i, z_k}$$

3 实例分析

选取流感病人信息系统来具体解释属性值信息系统中属性变化对于流图的影响, 此信息系统也是粗糙集理论中 Pawlak 及其同行广泛用于数据分析的经典数据表格 (见表 1 至表 3^[14]), 可作为上述定理的具体应用验证. 表 1 中属性 Headache 和 Temperature 作为条件属性, 属性 Flu 作为决策属性, 全域 U 由六位病人组成, 即 $U = \{p1, \dots, p6\}$. 表 2 中属性 Muscle-pain 和 Temperature 作为条件属性, 属性 Flu 作为决策属性, 全域 U 包含五个对象, 共有六位病人即 $\{p1, \dots, p6\}$. 为表述方便, 用各属性英文名称的首字母的大写斜体形式表示各属性, 相应的首字母的小写斜体形式表示属性值, 其值域为二值 $\{1 = \text{yes}, 0 = \text{no}\}$ 或三值 $\{1 = \text{very high}, 2 = \text{high}, 3 = \text{normal}\}$.

表 1 中的变量分布记作 P_1 , 此系统中存在流图 $h_0 \rightarrow t_3 \rightarrow f_0$, 因为:

$$[x]_{h_0, f_0} = \{p4\} = [x]_{t_3, h_0}$$

$$|[x]_{t_3}| = |[x]_{t_3, h_0}|$$

表 1 流感病人信息系统 1

Table 1 The first information system about patients suffering from flu

Patient	Headache(<i>H</i>)	Temperature(<i>T</i>)	Flu(<i>F</i>)
<i>p1</i>	no	high	yes
<i>p2</i>	yes	high	yes
<i>p3</i>	yes	very high	yes
<i>p4</i>	no	normal	no
<i>p5</i>	yes	high	no
<i>p6</i>	no	very high	yes

表 2 流感病人信息系统 2

Table 2 The second information system about patients suffering from flu

Fact no.	Muscle-pain(<i>M</i>)	Temperature (<i>T</i>)	Flu (<i>F</i>)	Patient name
1	yes	high	yes	<i>p1</i>
2	no	high	yes	<i>p2</i>
3	yes	very high	yes	<i>p3, p6</i>
4	yes	normal	no	<i>p4</i>
5	no	high	no	<i>p5</i>

表 3 添加新属性 Muscle-pain(or Headache)得到的新数据表格

Table 3 Data table obtained from Table 1(or Table 2) by adding the attribute Muscle-pain(or Headache)

Patient	Headache (<i>H</i>)	Muscle-pain(<i>M</i>)	Temperature (<i>T</i>)	Flu (<i>F</i>)
<i>p1</i>	no	yes	high	yes
<i>p2</i>	yes	no	high	yes
<i>p3</i>	yes	yes	very high	yes
<i>p4</i>	no	yes	normal	no
<i>p5</i>	yes	no	high	no
<i>p6</i>	no	yes	very high	yes

$$|[x]_{t_3, f_0}| = |[x]_{f_0, t_3, h_0}|$$

$$P_1(f_0 | t_3, h_0) = P_1(f_0 | t_3) = 1$$

表 2 中的变量分布记作 P_2 , 此系统中存在流图 $m_1 \rightarrow t_3 \rightarrow f_0$, 因为:

$$[x]_{m_1, f_0} = \{p4\} = [x]_{t_3, m_1}$$

$$|[x]_{t_3}| = |[x]_{t_3, m_1}|$$

$$|[x]_{t_3, f_0}| = |[x]_{f_0, t_3, m_1}|$$

$$P_2(f_0|t_3, m_1) = P_2(f_0|t_3) = 1$$

对于表1,若继续有新属性 Muscle-pain 被观测到,添加新属性后所形成的新信息系统为表3;对于表2,添加新属性 Headache 后也形成新信息系统,也是表3. 令表3中的变量分布记作 P .

显然,属性 Muscle-pain 未影响表1中属性 Headache, Temperature, Flu 以及三者的任意组合诱导的全域的划分. 或者说,属性 Headache, Temperature, Flu 没有受到属性 Muscle-pain 的影响. 但是表2中的属性 Muscle-pain, Temperature, Flu 皆受到新属性 Headache 的影响.

关于表3,通过概率计算有:

$$[x]_{h_0, f_0} = \{p4\}$$

$$[x]_{m_1, f_0} = \{p4\}$$

$$P(f_0|t_3, h_0, m_1) = P(f_0|t_3) = 1 = P(f_0|t_3, m_1)$$

$$P(t_3|h_0, m_1) = P(t_3|h_0) = 1/3$$

$$P(t_3|h_0, m_1) \neq P(t_3|m_1)$$

针对表1,依据定理1可知新系统中流图 $h_0 \rightarrow t_3 \rightarrow f_0$ 仍成立,同时还存在新流图 $m_1 \rightarrow h_0 \rightarrow t_3 \rightarrow f_0$. 针对表2,由定理2可知流图 $m_1 \rightarrow t_3 \rightarrow f_0$ 在新系统中仍成立.

4 总 结

本文主要探究了属性-值形式的信息系统中所抽取的流图对于添加新属性所导致的原始系统变化而产生迁移及变化的问题. 一个信息系统对应一个概率分布. 新属性的增加,必然引起系统中对象域的划分发生变化,系统所对应的概率分布将受影响. 通过对属性诱导的等价类的相关计算,研究发现,若旧系统中属性间的 Markov 性质在新系统中保持不变,则可得旧系统中所抽取的这些属性间的流图为新系统中这些属性间的流图或流图中的基本构件. 若进

一步关于由输入属性和输出属性诱导的等价类中的每个成员皆满足 Markov 性质,则旧系统中所抽取的这些属性间的流图即为新系统中这些属性间的流图. 此外,文中还给出了新属性对于流图的形状变化的细节研究. 伴随新证据的习得,人们会重新审视以前的结论并做相应的更新,此为人类学习和思维中的一种常见现象. 流图关于新属性的变化为理解和刻画这一现象提供新的理论依据,同时这种变化也可被用于人类行为的预测^[15]、属性约简^[16-17]和增量学习^[18]等相关研究.

参考文献

- [1] Pawlak Z. Rough sets. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.
- [2] Frege G. Grundgesetzen der Arithmetik. Volume 2. Jena: Verlag von Herman Pohle, 1903, 69.
- [3] Pawlak Z, Skowron A. Rudiments of rough sets. Information Sciences, 2007, 177(1): 3-27.
- [4] Pawlak Z. In pursuit of patterns in data reasoning from data-the rough set way//The 3rd International Conference on Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2002: 1-9.
- [5] Ford L R, Fulkerson D R. Flows in networks. Princeton: Princeton University Press, 1973, 1-35.
- [6] Yao N, Miao D Q. Identification of structures and causation in flow graphs. Information Sciences, 2019, 486: 287-309.
- [7] Pawlak Z. Decision trees and flow graphs//International Conference on Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2006: 1-11.
- [8] Butz C J, Yan W, Yang B. An efficient algorithm for inference in rough set flow graphs//Peters J F, Skowron A. Transactions on Rough Sets V. Springer Berlin Heidelberg, 2006: 102-122.
- [9] Chitcharoen D, Pattaraintakorn P. Novel matrix forms of rough set flow graphs with applications to

- data integration. Computers & Mathematics with Applications, 2010, 60(10): 2880—2897.
- [10] Lewicki A, Eberbach E. Learning network flow based on rough set flow graphs and ACO clustering in distributed cognitive environments// International Workshop on Software Engineering for Cognitive Services. New York, NY, USA: ACM, 2018: 18—24.
- [11] 姚宁, 苗夺谦, 张志飞. 因果信息在不同粒度上的迁移性. 计算机科学, 2019, 46(2): 178—186. (Yao N, Miao D Q, Zhang Z F. Transportability of causal information across different granularities. Computer Science, 2019, 46(2): 178—186.)
- [12] Liao S J, Zhu Q X, Qian Y H, et al. Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs. Knowledge-Based Systems, 2018, 158: 25—42.
- [13] Pawlak Z. Rough sets: theoretical aspects of reasoning about data. Springer Berlin Heidelberg, 1991, 2—4.
- [14] Pawlak Z. Some issues on rough sets// Transactions on Rough Sets I. Springer Berlin Heidelberg, 2004: 1—58.
- [15] Cao L B. In-depth behavior understanding and use: The behavior informatics approach. Information Sciences, 2010, 180(17): 3067—3085.
- [16] Dai J H, Hu Q H, Hu H, et al. Neighbor inconsistent pair selection for attribute reduction by rough set approach. IEEE Transactions on Fuzzy Systems, 2018, 26(2): 937—950.
- [17] Tan A H, Wu W Z, Qian Y H, et al. Intuitionistic fuzzy rough set-based granular structures and attribute subset selection. IEEE Transactions on Fuzzy Systems, 2019, 27(3): 527—539.
- [18] Xu J F, Miao D Q, Zhang Y J, et al. A three-way decisions model with probabilistic rough sets for stream computing. International Journal of Approximate Reasoning, 2017, 88: 1—22.

(责任编辑 杨可盛)