

基于概念格的非数值型数据聚类稳定性分析

智慧来

(河南理工大学计算机科学与技术学院 焦作 454000)

摘要 概念稳定性的计算是一个 NP-完全问题,且概念稳定性在聚类分析中具有十分重要的意义。概念格自身特点决定了其适合作为概念稳定性分析的工具。为了获得一个形式化的计算概念稳定性的方法,定义了概念的核心对象集,并给出了计算核心对象集的方法,进而得到了计算概念稳定性的方法。另外,根据概念格的对偶原理,得到了一个计算类标签的方法。最后用例子说明概念稳定性在聚类分析中的应用。

关键词 形式概念分析,概念格,聚类稳定性,核心对象集

中图分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.051

Clustering Stability Analysis for Non-numeric Data Based on Concept Lattice

ZHI Hui-lai

(School of Computer Science and Technology, He'nan Polytechnic University, Jiaozuo 454000, China)

Abstract Stable concepts usually represent strong correlation with real world entities and the calculation of concept stability, which is proven as an NP-complete problem, plays an important role in clustering analysis. To precisely calculate concept stability, concept lattice was used as the analysis model. At first, the definition of kernel object set as well as the way to find the kernel object set of a concept was proposed, and then concept stability was calculated based on kernel object set. Meanwhile, the method of calculating kernel attribute set of a given concept was derived directly based on the principle of duality of concept lattice. At last, an example was given to illustrate the application of concept stability.

Keywords Formal concept analysis, Concept lattice, Clustering stability, Kernel object set

1 引言

数据聚类是数据统计与分析中的重要技术,早在 20 世纪 80 年代就已经在各个领域广泛应用^[1]。通过聚类分析,可以将共同特征从一群事物中提取出来,并形成反映这些事物共同特性的思维单位,亦即概念。这一群事物构成了概念的外延,其共同特征构成了概念的内涵。概念稳定性体现了当外延或内涵缩小时概念保持不变的可能性。若一个概念的稳定性很强,则表明其外延中的对象具有极高的关联性^[2]。概念稳定性已经在分类树的建立和简化中得到应用^[3,4],并已经证明度量一个概念的稳定性是一个 NP-complete 问题^[5]。在本文中,将以非数值型数据为研究对象,提出计算概念稳定性的方法。

研究表明数值型对象聚类方法并不适合非数值型对象聚类^[6,7]。在数值型聚类分析中最常用的测距方式是使用曼哈顿距离,它相当于每个分量的绝对差值之和。在得到距离值之后,对象被联系起来,并通过分离和融合构建一个分类结构。

若对象 a 和 b 分别用向量表示为 $a=(a_1, a_2, \dots, a_n)$, $b=(b_1, b_2, \dots, b_n)$, 则曼哈顿距离的基本形式是:

$$(1) d(a, b) = |a_1 - b_1| + |a_2 - b_2| + \dots + |a_n - b_n|$$

$$(2) d(a, b) = u_1 |a_1 - b_1| + u_2 |a_2 - b_2| + \dots + u_n |a_n - b_n|,$$

$$u_1 + u_2 + \dots + u_n = 1$$

曼哈顿距离的第一种形式不考虑向量中分量间的关系,各个分量互不关联且重要性相同。曼哈顿距离的第二种形式考虑了分量的权重,即各个分量的重要性。分量加权的方式在模糊聚类^[8]、高维度数据聚类^[9]、混合数据类型聚类^[10]中频繁使用,计算分量权重的方法也得到了改进。

曼哈顿距离不考虑分量之间的关系,而实际上各个分量往往都不是互相独立的,分量之间存在着大量的一对一或者多对一的蕴含关系^[11]。假如用 $a, b, c, \dots$ 等小写字母表示分量,则分量之间的蕴含关系可以表示为:

$$a \rightarrow b, c \rightarrow d, \dots$$

$$ac \rightarrow b, acd \rightarrow e, \dots$$

由于采用曼哈顿距离没有考虑到分量之间的蕴含关系,距离的计算势必会出现偏差。

例 1:考虑鱼、蛙、狗 3 种动物:鱼是生活在水中、能运动且有四肢的动物;蛙是既能生活在水中又能生活在陆地、能运动且有四肢的动物;狗是生活在陆地、能运动、有四肢且能哺乳的动物。这里把鱼的腹鳍看作是四肢,表 1 中的“生物与水”形式背景来源于文献^[13],同样也把鱼的腹鳍看做是四肢。

若采用曼哈顿距离,容易得到 $d(\text{鱼}, \text{蛙}) = d(\text{蛙}, \text{狗})$,这显然和我们的常识不符合(鱼和蛙的相似性远大于蛙和狗的

相似性)。

这是因为在水中、能运动、有四肢、生活在陆地与能哺乳这些属性并不是互相独立的,而是存在着蕴含关系,比如对于任何一种动物:

如果有四肢,则一定蕴含能运动;

如果能生活在陆地也能生活在水中,则一定蕴含有四肢。

认知科学中的原型理论是研究分级归类模式的理论,它将范畴解释为由一些不同状态元素所构成的集合,并衍生出范畴内层级的概念^[12]。形式概念分析是研究知识表示的领域,与原型理论有许多共同之处,它的核心数据结构概念格是对概念以及概念之间关系的描述,由于它便于概念结构的开发和讨论,因此在某种意义上它已经变成了一种外部认知的手段^[13,14]。

概念格蕴含了对象之间以及属性之间的关联关系^[15],能够克服曼哈顿距离的缺陷。基于上述考虑,本文将利用概念格作为度量距离以及聚类稳定性分析的工具,提出计算聚类稳定性的方法,并基于概念格构造对象的分类结构。

2 概念格的基本概念

形式背景与形式概念是形式概念分析的两个柱石。下面将逐一给出形式概念分析的基本概念,这些内容均出自于Ganter R的著作“Formal Concept Analysis”^[13],但表述形式有所不同。

定义 1 一个形式背景是一个三元组 $K=(G, M, I)$, G 和 M 是两个集合, $I \subseteq G \times M$ 是一个二元关系,我们将 G 的元素称为对象, M 的元素称为属性, I 表示对象和属性之间的关系, gIm 或 $(g, m) \in I$ 表示对象 g 具有属性 m 。

定义 2 在形式背景 $K=(G, M, I)$ 中, $A \subseteq G$, 定义 $f(A) = \{m \in M | gIm, \forall g \in A\}$ (A 中对象共同属性的集合);相应地, $B \subseteq M$, 定义 $g(B) = \{g \in G | gIm, \forall m \in B\}$ (具有 B 中所有属性的对象的集合)。

定义 3 在形式背景 $K=(G, M, I)$ 中, $A \subseteq G, B \subseteq M$, 如果有 $f(A)=B, g(B)=A$, 则称二元组 $C=(A, B)$ 是一个概念, A 是概念 C 的外延, B 是概念 C 的内涵, 并记 K 的全体概念为 $B(K)$ 。

定义 4 在形式背景 $K=(G, M, I)$ 中, 概念 $C_1=(A_1, B_1), C_2=(A_2, B_2) \in B(K)$, 定义 $C_1 \leq C_2 \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$, 并称 C_2 是 C_1 的父概念, C_1 是 C_2 的子概念, $\leq$ 称为概念的层次序, $B(K)$ 中的所有概念用这种序组成的集合称为概念格, 记作 $L(G, M, I)$ 。

定义 5 一个形式背景 (G, M, I) 称为净化的, 如果任意两个满足 $f(g)=f(h)$ 的元素 $g, h \in G$ 都有 $g=h$, 而且对偶地任意两个满足 $g(m)=g(n)$ 的元素 $m, n \in M$ 都有 $m=n$ 。

定义 6 一个净化的形式背景 (G, M, I) 称为纯化的, 如果满足对任意一个元素 $g \in G$ 都不存在 $A \subseteq G$ 使得 $\bigcap_{a \in A} f(a) = f(g)$ 成立, 而且对偶地满足对任意一个元素 $m \in M$ 都不存在 $B \subseteq M$ 使得 $\bigcap_{b \in B} g(b) = g(m)$ 成立。

根据纯化形式背景的定义可知, 表 1 中的形式背景不是纯化形式背景, 因为存在 $g(a)=g(b)$ 和 $g(a) \cap g(b) \cap g(d) = g(e)$ 。表 2 中的形式背景是一个纯化形式背景。

表 1 一个非纯化的形式背景

	a	b	c	d	e
1	*	*		*	*
2	*	*	*		
3				*	
4	*	*	*		

表 2 一个纯化形式背景

	a	b	c	d
1	*			*
2		*	*	
3	*	*	*	
4	*	*		*

定义 7 在概念格 $L(G, M, I)$ 中, 如果一个概念具有形式 $(g(f(g)), f(g))$ 且 $g \in G$, 则称 $(g(f(g)), f(g))$ 是一个对象概念, 并称 $(g(f(g)), f(g))$ 是对象 g 所在的概念; 若 $A \in G$ 则称 $(g(f(A)), f(A))$ 是对象集 A 所在的概念, 若从分类的角度则称 $(g(f(A)), f(A))$ 是类 A 所在的概念。

一个对象概念的例子: 在表 2 中的形式背景中, 存在一个概念 $(\{1, 4\}, \{a, d\})$, 这个概念是一个对象概念, 因为这个概念可以表示为 $(g(f(1)), f(1))$ 。

3 基于概念格的聚类稳定性分析

在概念格中, 每一个概念都对应了一个分类。对于不同的类, 对象之间的紧密程度不同, 类的稳定性也不同。从直观上看, 如果在类中去掉一个或若干个对象后, 若这个类保持不变, 则这个类的稳定性相对较强; 反之, 这个类的稳定性相对较弱。根据这一思路, 可以得到概念稳定性的度量公式, 见定义 8。在原型理论中, 一个类的典型对象具有这个类的典型特征, 受此启发, 我们定义了一个概念的核心对象集, 见定义 9。

定义 8^[16] 对于概念格 $L(G, M, I)$ 中的一个概念 $C=(A, B)$, 其稳定性为 $s(C) = |\{D \subseteq A | f(D)=B\}| / 2^{|A|}$ 。

定义 9 对于给定的概念 (A, B) , 如果对象集合 R 满足 $f(R)=B$, 且对于任意的 $R_1 \subset R$ 有 $f(R_1) \supset f(R)$, 则称 R 是概念 (A, B) 的核心对象集。

核心对象集是这个概念外延中全体概念的一个代表, 它具有这个概念的全部属性特征。另外, 值得注意的是, 与信息系统的属性约简类似, 一个概念的核心对象集不一定是唯一的, 可能存在若干个核心对象集。

定理 1 在纯化形式背景中, 对象概念是上确界不可约元, 上确界不可约元一定是对象概念。

证明: 首先证明在纯化形式背景中对象概念一定是上确界不可约元, 然后证明上确界不可约元一定是对象概念。

(1) 证明纯化形式背景中对象概念一定是上确界不可约元。反证法, 假设对象概念 (A, B) 不是上确界不可约元, 即对象概念有两个或两个以上的下近邻, 其中的两个下近邻记作 $(A_t, B_t), t$ 为指标集, $t \in T$ 。由于 (A, B) 是对象概念, 因此存在对象 $g \in A$ 使得 $f(g)=B$ 。根据概念格基本定理, 有 $B = \bigcap_{t \in T} B_t$, 又因为 $f(A_t) = B_t (t \in T)$, 则有 $f(g) = \bigcap_{t \in T} f(A_t)$ ($t \in T$), 因此 g 是可约简的, 与纯化形式背景的前提矛盾, 故定理成立。

(2) 证明纯化形式背景中上确界不可约元一定是对象概念。反证法, 令概念 (A, B) 是一个上确界不可约元, 假设 (A, B) 不是对象概念, 则对于其概念外延中的对象 $g_1, g_2, \dots$,

$g_{|A|}$ 不存在一个 $1, 2, \dots, |A|$ 的排列 $p_1, p_2, \dots, p_{|A|}$ 使得 $f(g_{p_1}) \supseteq f(g_{p_2}) \supseteq \dots \supseteq f(g_{p_{|A|}})$ 成立, 这意味着至少存在两个对象 g_i, g_j , 使得 $f(g_i) \supseteq f(g_j)$ 和 $f(g_j) \supseteq f(g_i)$ 都不成立, 则必有 $f(g_i) \cap f(g_j) = f(A)$ 成立, 这意味着 (A, B) 至少有两个子概念, 其中一个概念的外延中包含 g_i , 另一个概念的外延中包含 g_j , 这显然与上确界不可约元只有一个下近邻矛盾, 故定理成立。

值得注意的是, 若一个形式背景不是纯化形式背景, 那么对象概念和上确界不可约元没有必然的联系, 也就是说此时对象概念不一定是上确界不可约元。

定理 2 若某个概念是一个对象概念, 则其核心对象集是这个概念的标签对象。

定理 3 若某个概念不是一个对象概念, 则它的一个核心对象集为任意两个子概念的核心对象集的并集。

证明: 令 $P = (A, B)$ 为一个非对象概念, P 有子概念 $P_1 = (A_1, B_1), P_2 = (A_2, B_2), \dots, P_n = (A_n, B_n)$, 这些子概念的核心对象集分别为 $C_1, C_2, \dots, C_n$, 那么根据概念格的基本定理有 $A = \bigcup_{i=1}^n A_i$. 又 $f(A) = \bigcap_{i=1}^n f(C_i)$, 当 $i \neq j$ 时 $f(C_i) \cap f(C_j) \supseteq \bigcap_{i=1}^n f(A_i) = f(A)$, 现在证明仅等号成立。(反证法) 假设 “ $\supseteq$ ” 成立, 令 $B' = f(C_i) \cap f(C_j)$, 那么存在一个概念 $(g(B'), B')$ 使得 $(A, B) > (g(B'), B') > P_i (i=1, 2, \dots, n)$, 这与 (A, B) 是 $P_i (i=1, 2, \dots, n)$ 的父概念矛盾。所以有当 $i \neq j$ 时, $f(C_i) \cap f(C_j) = f(A)$, P 的一个核心对象集为 $C_i \cup C_j (i \neq j, i, j \in 1, 2, \dots, n)$, 定理成立。

由定理 3 可知, 若一个非对象概念有若干子概念, 通过任意两个子概念核心对象的并集可以得到一个核心对象集。

概念的核心对象集的递归计算公式:

(1) 若一个概念是对象概念, 则其核心对象集是这个概念的标签对象;

(2) 若一个概念不是对象概念, 则其核心对象集是其任意两个子概念核心对象集的并集。

在概念格上, 从其中的最小概念开始层次遍历每一个概念, 在遍历概念的同时使用上述递归公式就可以得到所有概念的核心对象集。

定理 4 若一个类 C 对应的概念为 (A, B) , R 是概念 (A, B) 唯一的核对象集, 则类 C 的稳定性为 $s(C) = 2^{-|R|}$ 。

证明: R 是概念 (A, B) 的核对象集, 那么对于任意的 $R' \supseteq R$, 都有 $f(R') = B$. R' 的取值范围是 $[R, A]$, 在这个区间里, 存在 $2^{|A|-|R|}$ 个这样的对象集合, 根据类的稳定性的定义, $s(C) = 2^{|A|-|R|} / 2^{|A|} = 2^{-|R|}$, 定理成立。

定理 5 若一个类 C 对应的概念为 (A, B) , 并且 (A, B) 有 n 个核心对象集, 分别为 $R_1, R_2, \dots, R_n$, 则类 C 的稳定性

$$s(C) = \left(\sum_{i=1}^n 2^{|A|-|R_i|} - \sum_{1 \leq i < j \leq n} 2^{|A|-|R_i \cup R_j|} + \sum_{1 \leq i < j < k \leq n} 2^{|A|-|R_i \cup R_j \cup R_k|} + \dots + (-1)^{n+1} 2^{|A|-|R_1 \cup R_2 \cup \dots \cup R_n|} \right) / 2^{|A|}.$$

证明: $R_1, R_2, \dots, R_n$ 是概念 (A, B) 的核对象集, 那么对于任意的 $A \supseteq R'_i \supseteq R_i$ 都有 $f(R'_i) = B$ 易知在区间 $[R_i, A]$ 中有 $2^{|A|-|R_i|}$ 个对象集。又因为在区间 $[R_i \cup R_j, A]$ 中有 $2^{|A|-|R_i \cup R_j|}$ 个对象集合, 在区间 $[R_i \cup R_j \cup R_k, A]$ 中有 $2^{|A|-|R_i \cup R_j \cup R_k|}$ 个对象集, 在区间 $[R_1 \cup R_2 \cup \dots \cup R_n, A]$ 中有 $2^{|A|-|R_1 \cup R_2 \cup \dots \cup R_n|}$ 个对

象集, 根据包含排除原理的推广形式可知, 一共有 $\sum_{i=1}^n 2^{|A|-|R_i|} - \sum_{1 \leq i < j \leq n} 2^{|A|-|R_i \cup R_j|} + \sum_{1 \leq i < j < k \leq n} 2^{|A|-|R_i \cup R_j \cup R_k|} + \dots + (-1)^{n+1} 2^{|A|-|R_1 \cup R_2 \cup \dots \cup R_n|}$ 个这样的集合 D , 使得 $D \subseteq A, f(D) = B$ 。

根据类的稳定性的定义可知定理成立。

推论 1 若一个类 C 对应的概念为 (A, B) , (A, B) 有两个核心对象集 R_1 和 R_2 , 则类 C 的稳定性 $s(C) = (2^{|A|-|R_1|} + 2^{|A|-|R_2|} - 2^{|A|-|R_1 \cup R_2|}) / 2^{|A|}$ 。

4 基于概念格的类标签计算

一个类的标签由这个类所在概念的内涵确定。内涵中的属性并不是互相独立的, 属性之间存在蕴含关系, 因此存在冗余的属性, 可以对这些冗余进行约简。

根据概念格的对偶原理, 可以得到下列定义和定理, 对定理不再进行证明。

定义 10 在一个概念格中, 如果一个概念具有形式 $(g(m), f(g(m)))$ 且 $m \in M$, 则称 $(g(m), f(g(m)))$ 是一个属性概念。

定义 11 对于给定的概念 (A, B) , 如果属性集合 R 满足 $g(R) = A$, 且对于任意的 $R_1 \subset R$ 有 $g(R_1) \supsetneq g(R)$, 则称 R 是概念 (A, B) 的核属性集。

定理 6 在纯化形式背景中, 属性概念是下确界不可约元, 下确界不可约元一定是属性概念。

定理 7 若某个概念是一个属性概念, 则其核属性集是这个概念的标签属性。

定理 8 若某个概念不是一个属性概念, 则它的一个核属性集为其任意两个父概念的核属性集的并集。

5 算法分析与实例研究

在文献[5]中, Kuznetsov S. O 提出了一个计算概念稳定的算法, 其主要思想是: 对于一个给定的概念 (A, B) , 初始化其稳定性为 0, 然后从小到大依次搜索 A 的所有子集 $X (X \subseteq A)$ 并进行判断, 若 $f(X) = B$, 则增加这个概念的稳定性, 否则搜索下一个子集。

此外, Kuznetsov S. O 还对提出的算法进行了分析, 指出: 对于形式背景 (G, M, I) 中的概念 (A, B) , 若令 $|\{D \subseteq A | f(D) = B\}| = k$, 则算法的复杂度为 $O(|G|^3 |M| k)$ 。

本文的研究是基于概念格模型的, 在应用中首先需要考虑的是建立概念格的效率。Sergei O. Kuznetsov 和 Sergei Obiedkov 研究了 10 种最具有代表性的概念格构造算法, 得到的结论是: 对于一个形式背景 (G, M, I) , 若采用插入对象的方式建立概念格, 其时间复杂度是 $O(|G|^2 |M| L)$, L 表示概念格中概念的数量^[17]。在计算一个概念的核心对象集时, 只需要对这个概念的子概念进行递归搜索, 与搜索 A 的所有子集相比, 显然缩小了搜索范围。对于概念格中的任何一个概念, 显然其子概念的数量小于 L , 因此本文计算方法的时间复杂度为 $O(|G|^2 |M| L^2)$ 。

形式背景 (G, M, I) 的概念数量 L 不仅和 G, M 有关, 而且和关系 I 有关, 难以定量分析, 因此可以采用实验的方法进行分析。随机生成形式背景, 属性个数固定为 20 个, 对象属性间存在关系的概率为 0.2、0.25、0.33。图 1 绘出了概念总

数随着对象数量增加时的关系曲线(L1 对应的概率为 0.2, L2 对应的概率为 0.25, L3 对应的概率为 0.33)。

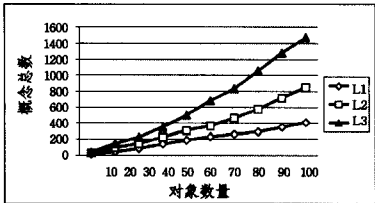


图1 不同概率下概念格规模的变化曲线

从实验结果可以看出,概念格的规模随着对象的增加而增大, L 与 $|G|$ 显然不是指数关系, L 远小于 $2^{|G|}$ 。由实验结果可估算,假如一个概念的外延中有 100 个对象,按照 Kuznetsov S. O 的方法则需要搜索的子集有 2^{100} 个(2^{100} 约等于 1.2676506×10^{30})。而利用概念格,当对象属性间存在关系的概率为 0.33 时,搜索的概念数量在 1500 个左右;当对象属性间存在关系的概率下降到 0.2 时,搜索的概念数量下降到 400 个左右。显然,本文的方法缩小了搜索范围,能够提高计算效率。

例 2:在匈牙利的一个科教电影“生物和水”中,存在 8 种生物,分别是蚂蝗(1)、鱼(2)、蛙(3)、狗(4)、水草(5)、芦苇(6)、豆(7)、玉米(8)。同时,这个电影强调了 9 种生物属性,即 a(需要水)、b(在水中生活)、c(在陆地生活)、d(有叶绿素)、e(双子叶)、f(单子叶)、g(能运动)、h(有四肢)、i(哺乳)。这些生物构成的形式背景见表 3,其概念格见图 2。

表3 “生物和水”形式背景

	a	b	c	d	e	f	g	h	i
1	*	*					*		
2	*	*					*	*	
3	*	*	*				*	*	
4	*		*				*	*	*
5	*	*		*		*			
6	*	*	*	*		*			
7	*		*	*	*				
8	*		*	*		*			

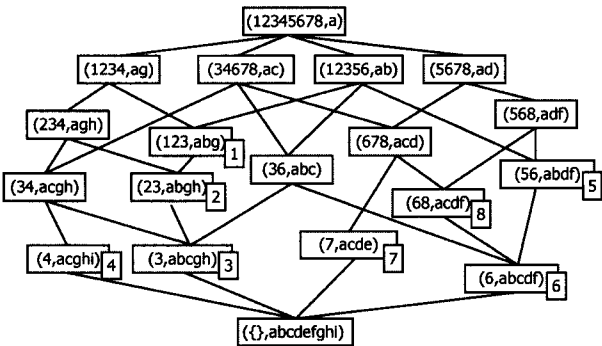


图2 “生物和水”形式背景的概念格

判定一个对象与一个类的从属关系时,稳定性是聚类分析中的一条重要准则。在例 2 中,蛙和狗形成一个类 A_1 (能生活在陆地,能运动,有四肢的生物),蛙和鱼形成一个类 A_2 (能生活在水中、能运动、有四肢的生物)。在此粒度下,若蛙只能属于一个类,则需要分别计算类 A_1 和类 A_2 的稳定性:

类 A_1 所在概念为 (34,acgh), $s(A_1) = s((34,acgh)) = 0.25$;
类 A_2 所在概念为 (23,abgh), $s(A_2) = s((23,abgh)) = 0.5$ 。

比较两个类的稳定性,蛙和鱼这一类的稳定性更高,应将蛙和鱼为归为一类,而不是将蛙和狗归为一类。

根据概念格很容易建立对象的分类树,见图 3。每个内部结点有两个部分,即类的编号和这个类的标签。

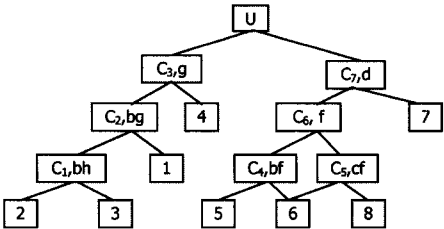


图3 “生物和水”形式背景中对象的分类树

显然,上述结果分类层次清晰,且符合生物学中支序分类学的观点^[18]。

结束语 考虑到非数值对象的特点以及曼哈顿距离的不足之处,本文利用概念格作为度量对象间距以及聚类分析的工具。本文最重要的工作是提出了核心对象的概念,以核心对象为基础找到了计算概念稳定性的方法;然后,根据概念格的对偶原理,得到了类标签的计算方法。

本文的研究对象是非数值型同构数据,但在实际应用中往往存在大量的异构数据(如表 4)。在下一步研究中,我们将本文的思路扩展到异构数据,建立面向异构数据的概念格模型,提出混合型异构数据的聚类稳定性分析方法。

表4 形式背景 K

	A	O	T	V
g ₁	e ₁ , e ₂	a, b	[t ₁ , t ₂]	[62, 73]
g ₂	e ₁ , e ₃	d, e	t ₁	[30, 73]
g ₃	e ₂	a, c	t ₂	[20, 35]
g ₄	e ₂ , e ₄	a	[t ₁ , t ₂]	[40, 50]

参 考 文 献

[1] Jain A K, Dubes R C. Algorithms for Clustering Data[M]. Prentice Hall, 1988

[2] Melo C, Le-Grand B, Aufaure M A, et al. Extracting and visualising tree-like structures from concept lattices[C]// 15th International Conference on Information Visualisation. London, United Kingdom, IEEE Computer Society, 2011: 261-266

[3] Brito P, Polaillon G. Homogeneity and stability in conceptual analysis[C]// Amedeo Napoli, Vilem Vychodil, eds. CLA, 2011: 251-263

[4] Kuznetsov S, Obiedkov S, Roth C. Reducing the representation complexity of lattice-based taxonomies[J]. Conceptual Structures: Knowledge Architectures for Smart Applications Lecture Notes in Computer Science, 2007, 4604: 241-254

[5] Kuznetsov S O. On stability of a formal concept [J]. Annals of Mathematics and Artificial Intelligence, 2007, 49(1-4): 101-115

[6] Gan G, Wu J. Subspace clustering for high dimensional categorical data[J]. ACM SIGKDD Explorations Newsletters, 2004, 6(2): 87-94

[7] Zaki M, Peters M, Assent I, et al. CLICK: an effective algorithm for mining subspace clusters in categorical datasets[J]. Data and Knowledge Engineering, 2007(60): 51-70

[8] Tang Cheng-long, Wang Shi-gang, Xu Wei. New fuzzy c-means clustering model based on the data weighted approach[J]. Data

- [9] Bai Liang, Liang Ji-ye, Dang Chuang-yin, et al. A novel attribute weighting algorithm for clustering high-dimensional categorical data [J]. Pattern Recognition, 2011, 44(12): 2843-2861
- [10] Ji Jin-chao, Pang Wei, Zhou Chun-guang, et al. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data [J]. Knowledge-Based Systems, 2012, 30: 129-135
- [11] Wu Xin-dong, Zhang Cheng-qi, Zhang Shi-chao. Efficient mining of both positive and negative association rules[J]. ACM Transactions on Information Systems, 2004, 22(3): 381-405
- [12] Rosch E. Human categorization[C]// Warren N, ed. Studies in Cross-Cultural Psychology. Academic press, New York, 1977: 1-49
- [13] Ganter B, Wille R. Formal Concept Analysis: mathematical foundation [M]. New York: Springer-Verlag, 1999
- [14] Scaife M, Rogers Y. External cognition: how do graphical representations work[J]. International Journal of Human Computer Studies, 1996, 45: 185-213
- [15] 谢志鹏, 刘宗田. 概念格与关联规则发现[J]. 计算机研究与发展, 2000, 37(12): 1415-1421
- [16] Kuznetsov S O. Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity[J]. Nauchn. Tekh. Inf., Ser. 2 (Automat. Document. Math. Linguist.), 1990, 12: 21-29
- [17] Kuznetsov S O, Obiedkov S. Comparing Performance of Algorithms for Generating Concept Lattices[J]. Journal of Experimental and Theoretical Artificial Intelligence, 2002 (14): 189-216
- [18] Adams E N. Consensus techniques and the comparison of taxonomic trees[J]. Systematic Zoology, 1972, 21: 390-397

(上接第 224 页)

有些是很难通过我们给出的形式化框架进行定量或定性表示的, 比如行为的执行人员之间交换信息的次数。

但有些我们是可以通过逻辑推理得到一个定量或定性的指标评价值的, 比如并行执行的行为数, 由于并行执行的行为之间不应该有互相的输入输出关联关系, 而根据我们的定义, 流程白盒中的流程之间是无序与关系, 其中可能包含并行执行的行为, 因此, 生成目标匹配的业务流程片段后, 我们可以对流程片段的流程白盒中的各个流程包含的行为进行输入输出传递关系检查。若行为 A_1 在行为 A_2 的输入(输出)链上, 则称 A_1 和 A_2 是间接依赖的。通过对行为进行两两间接依赖关系检查, 可以得到那些不存在间接依赖关系的行为, 亦即可以并行执行的行为。

以上只以 3 种非功能目标为例, 简单探讨了对生成与客户目标匹配的业务流程片段评估的指标, 关于更深入的评估过程的形式化描述将作为我们后续的研究内容。

结束语 本文提出了一个基于目标的可重用的业务流程建模框架, 根据目标-行为分层关系以及细化后的行为属性, 自动构建分解后的所有目标对应的业务流程片段, 实现业务流程片段的重用; 并提出定性或定量地评估流程片段对功能目标的支持度, 在为用户进行目标-流程片段匹配时提供决策指导; 最后, 本文也简要探讨了如何形式化地评估生成的业务流程片段对非功能目标的支持度, 增加了用户对业务流程模型的评价维度。

基于本文的研究, 也提出了一些在未来工作中需要深入探讨的问题, 比如: 目标-行为分层关系中的功能目标一致性检查; 根据细化行为自动生成的流程片段集合后, 如何去除其中没有任何业务意义的业务流程片段; 如何根据行为对原目标的支持度自动推演出业务流程片段对相应目标的支持度等等。我们在未来的研究工作中将继续探索将其与其它研究领域的理论技术相结合, 如人工智能、流程挖掘等, 解决所提出的问题。

参考文献

- [1] Kueng P, Bichler P, Kawalek P, et al. How to compose an object-oriented business process model? [C]// Brinkkemper S, et al., eds. Method Engineering: Principles of Method Construction and Tool Support, Proceedings of the IFIP WG8. 1/WG8. 2 Working Conference, Atlanta, Chapman & Hall, London, August 1996: 94-110
- [2] Scherer E, Zölch M. Design of activities in shopfloor management: a holistic approach to organization at operational business levels in bpr projects[C]// Browne J, O'Sullivan D. eds. Re-engineering the Enterprise, Proceedings of the IFIP TC5/WG5. 7 Working Conference, 1995: 261-272
- [3] Jr J B, Persson A, Stima J. User guide of the knowledge management approach using enterprise knowledge patterns[R]. KTH, Sweden, 2001
- [4] Markovic I, Kowalkiewicz M. Linking Business Goals to Process Models in Semantic Business Process Modeling[C]// Proceedings of the 12th IEEE International EDOC Conference, Munich, Germany, September 2008
- [5] Yu E S K, Mylopoulos J. Understanding "why" in software process modeling, analysis, and design[C]// Proceedings of the 16th International Conference on Software Engineering. IEEE Computer Society Press, May 1994: 159-168
- [6] Soffer P, Wand Y. On the notion of soft-goals in business process modeling [J]. Business Process Management Journal, 2005, 11(6): 663-679
- [7] Markovic I, Pereira A C. Towards a Formal Framework for Re-use in Business Process Modeling[C]// Workshop on Advances in Semantics for Web services (semantics4ws), in conjunction with BPM '07. Brisbane, Australia, September 2007
- [8] Kueng P, Kawalek P. Goal-based business process models: creation and evaluation [J]. Business Process Management Journal, 1997, 3(1): 17-38
- [9] Hammer M, Champy J. Reengineering the Corporation: A Manifesto for Business Revolution[M]. New York: Harper Business, 1993
- [10] Ali R, Dalpiaz F, Giorgini P. A goal-based framework for contextual requirements modeling and analysis [J]. Requirements Engineering, 2010, 15: 439-458
- [11] Ould M. Business Processes: Modelling and Analysis for Re-engineering and Improvement [M]. Chichester: John Wiley & Sons, 1995
- [12] Cardoso E, Almeida J P A, Guizzardi R S S, et al. A method for eliciting goals for business process models based on non-functional requirements catalogues [J]. International Journal of Information System Modeling and Design (IJISMD), 2011, 2(2): 1-18