文章编号: 1003- 207(2011) 01- 0135- 07

动态信息系统中一种基于条件熵的核属性计算方法

梁德翠, 胡 培

(西南交通大学经济管理学院, 四川 成都 610031)

摘 要: 针对动态信息系统中新增和退出对象集的情况, 在信息观下提出了一种基于条件熵的核属性计算方法。先分别讨论仅有对象集进入、仅有对象集退出以及以上两种情形同时存在下条件熵的变化机制。由条件熵变化机制, 通过构造支持度矩阵和增量矩阵方式将新增或者退出的对象集一并参与比较计算, 快速求得条件熵, 进而获得核属性。最后, 由实例分析验证该方法的有效性。

关键词: 动态信息系统; 增量矩阵; 支持度矩阵; 条件熵; 核属性

中图分类号: TP18; C931 文献标识码: A

1 引言

一项决策问题, 决策者一般需解决两个基本问题: 一是在特定情况下, 如何做出好的决策; 二是对所做出的决策给出一种合理解释[1]。波兰数学家 Pawlak 于 1982 年提出的粗糙集理论[2] 正好为解决以上问题提供一种新的方法, 它在处理不精确性、模糊性及不确定性方面具有很好效果。粗糙集理论的基本思想是通过数据预处理、属性约简等过程, 最终获取决策规则。目前, 该理论已成功地应用于机器学习、知识获取、决策分析、知识发现、模式识别、专家系统和决策支持系统等领域[3]。

属性约简是粗糙集理论的核心问题[4], 也是知识获取的关键一步。属性约简已经证明是一个 NP 问题, 目前大多采用启发式方法求解, 其中一类重要的方法是基于核属性展开。关于求解核属性方面的研究已有很多[3, 5- 9, 11]。文献[5] 提出了基于差别矩阵的求核方法, 该方法可有效地减少计算量, 提高求核的效率, 但在不协调的决策表中得不到正确的核。文献[6] 在文献[5] 的差别矩阵定义基础上, 提出新的差别矩阵并证明了其求核方法的正确性, 但计算量太大。文献[7] 在文献[6] 基础上提出了新的改进的差别矩阵, 该方法可以得到代数观下的正确核属性, 也能够有效地降低计算复杂度。文献[8] 从代数观和信息观两方面对核属性进行了深入的探讨, 并给出了一种基于信息熵的核属性计算算法, 该算法可以有效地补充和完善文献[6] 中的算法。以上研究都集中在静态信息系统中, 在实际生活中系统是处于不断变化的, 既有对象集的进入, 也有现有对象集的退出[10, 11]。随着系统的动态变化, 已有的核属性可能不再有效, 需要对其进行动态修改。

在动态信息系统求核方面, 文献[12] 从代数观角度, 提出一种基于改进差别矩阵的核属性增量式更新算法, 该算法在更新差别时仅需插入某一行和某一列, 或删除某一行并修改相应的列, 因而提高了属性核的更新效率。文献[3] 从信息观角度, 分析了新增对象后条件熵的变化机制, 提出了一种基于条件熵的增量式核求解算法。以上的研究只考虑动态系统中出现新增对象集, 计算时按单个对象依次进行。对于动态系统中存在对象集退出和采用多对象同时比较计算的情况, 该方面研究较少。

而在实际动态系统中既有多个新增对象, 又有多个对象退出。为更加符合实际情况, 本文在文献[3] 基础上, 考虑系统中有新增或者退出对象集, 共分三种情形依次分析了对应条件熵的变化机制。在求解新的条件熵时借助增量矩阵和支持度矩阵, 将新增或退出的对象集一并参与比较计算, 实现动态系统下条件熵的快速求解。最后, 提出了一种基于条件熵的核属性计算方法。

收稿日期: 2010- 04- 29; 修订日期: 2011- 01- 10

基金项目: 国家自然科学基金资助项目(60873108)

作者简介: 梁德翠(1986-), 男(汉族), 江西瑞昌人, 西南交通大学经济管理学院, 博士研究生, 研究方向: 粗集决策.

2 基本概念

假设有一个信息系统由一个四元组 $S = < U, A, V, f >$,其中 U 是非空有限集合, 称为论域, 其中元素称为对象; A 是非空属性集, $V = \bigcup_{a \in A} V_a$, V_a 表示属性 a 取值构成的集合; $f: U \times A \rightarrow V$ 称为信息函数, 表示对于任意 $x \in U$, $a \in A$, $f(x, a) \in V_a$ 。若 $A = C \cup D$, $C \cap D = \phi$, 且 C 和 D 分别表示条件属性集和决策属性集, 则该信息系统称之为决策系统, 令 $D = \{d\}$ 。

定义 1[13]: 给定决策表 $S = < U, A, V, f >$, 其中 $A = C \cup D$, C 是条件属性集, D 是决策属性集, $U/C = \{X_1, X_2, \cdots, X_m\}$, $U/D = \{D_1, D_2, \cdots, D_n\}$, 决策表的条件熵定义为:

$$E(D \mid C) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\mid X_i \cap D_j \mid}{\mid U \mid} \frac{\mid D_j^C - X_i^C \mid}{\mid U \mid} \tag{1}$$

其中, X_i^C 表示 X_i 的补集, D_j^C 表示 D_j 的补集, $\mid \cdot \mid$ 表示对象集的基数。

定理 1[3]: 给定决策表 $S = < U, A, V, f >$, 其中 $A = C \cup D$, C 是条件属性集, D 是决策属性集。对于 $\forall a \in C$ 属于信息观下 D 的 C 核 $CORE_D(C)$ 的充分必要条件为:

$$E(D \mid C) < E(D \mid (C - \{a\})) \text{ 。} \tag{2}$$

根据定义 1 可得, 决策表的条件熵等价于:

$$E(D \mid C) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\mid X_i \cap D_j \mid}{\mid U \mid} \frac{\mid X_i - D_j \mid}{\mid U \mid} \tag{3}$$

为便于分析, 把各元素的公共部分 $1/\mid U \mid^2$ 提取出来, 令 $E(D_j \mid X_i) = \mid X_i \cap D_j \mid \mid X_i - D_j \mid$ 。那么决策表的条件熵为:

$$E(D \mid C) = \frac{1}{\mid U \mid^2} \sum_{i=1}^{m} \sum_{j=1}^{n} E(D_j \mid X_i) \tag{4}$$

从本质上来看公式(3)和(4)是相同的。

而根据公式(4)可把条件熵中的各元素一一对应到矩阵中, 矩阵结构如下:

$$\begin{bmatrix} E(D_1 \mid X_1) & E(D_2 \mid X_1) & \cdots & E(D_n \mid X_1) \\ E(D_1 \mid X_2) & E(D_2 \mid X_2) & \cdots & E(D_n \mid X_2) \\ \vdots & \vdots & \cdots & \vdots \\ E(D_1 \mid X_m) & E(D_2 \mid X_m) & \cdots & E(D_n \mid X_m) \end{bmatrix} \tag{5}$$

显然, 矩阵的行对应条件类, 而列则对应决策类。$E(D_j \mid X_i) = \mid X_i \mid \mid X_i \cap D_j \mid - \mid X_i \cap D_j \mid^2$。矩阵结构将便于对动态信息系统中条件熵的变化机制的理解和计算。

3 动态信息系统中条件熵的变化机制

在实际生活中信息系统处于动态变化的, 既有对象集的进入, 也有现有对象集的退出。给定一个决策表 $S = < U, A, V, f >$, $A = C \cup D$, C 是条件属性集, D 是决策属性集。约定: 定义两个时刻, t 和 $t+1$ 。其中, t 时刻为原有系统, $t+1$ 时刻为新增或者退出对象集后的系统。假设 t 时刻, 系统中 $U/C = \{X_1, X_2, \cdots, X_m\}$, $U/D = \{D_1, D_2, \cdots, D_n\}$, 此时条件熵记为 $E^{(t)}(D \mid C)$ 。下面分别对对象集的进入和退出情况进行讨论。

3.1 仅有对象集进入时条件熵的变化机制

文献[3]中已把新增的对象分成四种情况: (1) 不属于 U/C 的条件类, 也不属于 U/D 的决策类; (2) 不属于 U/C 的条件类, 但属于 U/D 的决策类; (3) 属于 U/C 的条件类, 但不属于 U/D 的决策类; (4) 属于 U/C 的条件类, 也属于 U/D 的决策类。本文借鉴文献[10, 11]构造矩阵方法, 将新增对象集的各种情况一并讨论。

假设 $t+1$ 时刻, 新增了 N 个对象集, 条件类中新增 l 个条件类, 而决策类中又新增 r 个决策类, 即 $U'/C = \{X'_1, X'_2, \cdots, X'_m, \cdots, X'_{m+l}\}$, $U'/D = \{D'_1, D'_2, \cdots, D'_n, \cdots, D'_{n+r}\}$, 其中记 U' 为新增对象集后系统的论域。$\forall X'_i \in U'/C$, 该条件类对应新增对象集数为 N_i , 若 $\forall D'_j \in U'/D$, 规则: $X'_i \rightarrow D'_j$ 中支持度新增对应的对象数为 N_{ij} , 示意图详见文献[10, 11]。

基于以上的假设条件, 按照公式(5), 先分析 $t+1$ 时刻矩阵中的元素:

$$E(D'_j \mid X'_i) = \mid X'_i \mid \mid X'_i \cap D'_j \mid - \mid X'_i \cap D'_j \mid^2$$

$$= (\mid X_i \mid + N_i)(\mid X_i \cap D_j \mid + N_{ij}) - (\mid X_i \cap D_j \mid + N_{ij})^2$$

$$= (\mid X_i \mid \mid X_i \cap D_j \mid - \mid X_i \cap D_j \mid^2) + (N \mid X_i \mid + N_i \mid X_i \cap D_j \mid + N_i N_{ij} - (N_{ij}^2 + 2N_{ij} \mid X_i \cap D_j \mid))$$

这里, $i = 1, 2, \cdots, m, \cdots, m+l$, $j = 1, 2, \cdots, n, \cdots, n+r$ 。

因此, $t+1$ 时刻条件熵 $E^{(t+1)}(D \mid C)$ 为:

$$E^{(t+1)}(D \mid C) = \frac{1}{(\mid U \mid + \sum_{i=1}^{m+l} N_i)^2} (\mid U \mid^2 E^{(t)}(D \mid C) + \sum_{i=1}^{m+l} ((N_i^2 + 2N_i \mid X_i \mid) - \sum_{j=1}^{n+r} (N_{ij}^2 + 2N_{ij} \mid X_i \cap D_j \mid)))$$

(6)

其中: $N_i = \sum_{j=1}^{n+r} N_{ij}$, $N = \sum_{i=1}^{m+l} N_i$, $|X_i| = \sum_{j=1}^{n} |X_i \cap D_j|$。

需要说明地是: 当 $i = m+1, m+2, \cdots, m+l$ 时, $|X_i| = 0$。当 $i = m+1, m+2, \cdots, m+l$ 或者 $j = n+1, n+2, \cdots, n+r$ 时, $|X_i \cap D_j| = 0$。

公式(6)中反映了 $E^{(t+1)}(D|C)$ 和 $E^{(t)}(D|C)$ 的关系, 式中 $E^{(t+1)}(D|C)$ 的修正除原有条件熵之外, 还需要考虑新增对象集分布在各条件类和决策类上的情况和原有系统中各条件类同决策类间的交集。因此, $t+1$ 时刻条件熵求解的关键在于 N_{ij} 和 $|X_i \cap D_j|$。对于 $\forall X_i \in U/C$, $\forall D_j \in U/D$, 规则: $X_i \rightarrow D_j$ 的支持度 $Supp(D_j|X_i) = |X_i \cap D_j|$。

由此借鉴文献[10, 11]的矩阵方法, 依次构造出原系统的支持度矩阵 $Supp^{(t)}(D|C)$ 和增量矩阵 $Inc^{(t+1)}(D|C)$:

$$Supp^{(t)}(D|C) = \begin{bmatrix} |X_1 \cap D_1| & |X_1 \cap D_2| & \cdots & |X_1 \cap D_n| \\ |X_2 \cap D_1| & |X_2 \cap D_2| & \cdots & |X_2 \cap D_n| \\ \vdots & \vdots & \cdots & \vdots \\ |X_m \cap D_1| & |X_m \cap D_2| & \cdots & |X_m \cap D_n| \end{bmatrix} \tag{7}$$

$$Inc^{(t+1)}(D|C) = \begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1n} & \cdots & N_{1(n+r)} \\ N_{21} & N_{22} & \cdots & N_{2n} & \cdots & N_{2(n+r)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ N_{m1} & N_{m2} & \cdots & N_{mn} & \cdots & N_{m(n+r)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ N_{(m+l)1} & N_{(m+l)2} & \cdots & N_{(m+l)n} & \cdots & N_{(m+l)(n+r)} \end{bmatrix} \tag{8}$$

矩阵的行对应条件类, 而列则对应决策类。公式(7)的支持度矩阵反映地是各条件类同各决策类间的交集情况, 而公式(8)的增量矩阵反映地是新增对象集在各条件类与决策类下的分布情况。

按照公式(6), 原有条件熵再加上公式(8)和(7), 可得 $t+1$ 时刻条件熵。在利用公式(7)和(8)计算条件熵时, 先通过增量矩阵的新增对象集, 再对应从支持度矩阵中找到交集。与文献[3]不同的是, 本文条件熵的求解不再是单个元素依次比较计算, 而是在进入系统前统一比较。

本文条件熵需计算: $(m+l)(n+r)(mc+n+1)$ 次。而利用文献[3]的算法最终条件熵需计算的次数除了与原有条件类与决策类有关, 还与进入对象集的个数, 有无新增条件类或决策类以及它们进入系统的先后顺序有关。

那么, $t+1$ 时刻支持度矩阵 $Supp^{(t+1)}(D|C)$ 为:

$$Supp^{(t+1)}(D|C) = Supp^{(t)}(D|C) + Inc^{(t+1)}(D|C)$$

$$= \begin{bmatrix} |X_1 \cap D_1| + N_{11} & |X_1 \cap D_2| + N_{12} & \cdots & |X_1 \cap D_n| + N_{1n} & \cdots & N_{1(n+r)} \\ |X_2 \cap D_1| + N_{21} & |X_2 \cap D_2| + N_{22} & \cdots & |X_2 \cap D_n| + N_{2n} & \cdots & N_{2(n+r)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ |X_m \cap D_1| + N_{m1} & |X_m \cap D_2| + N_{m2} & \cdots & |X_m \cap D_n| + N_{mn} & \cdots & N_{m(n+r)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ N_{(m+l)1} & N_{(m+l)2} & \cdots & N_{(m+l)n} & \cdots & N_{(m+l)(n+r)} \end{bmatrix} \tag{9}$$

此时为满足矩阵加法原则, 先对原有支持度矩阵进行以下处理: 对应新增条件类和决策类先增加 l 行, 增加 r 列并令各元素为 0。然后, 再同增量矩阵相加, 求得 $t+1$ 时刻支持度矩阵。由公式(9)可得, 矩阵中的元素为:

$$|X'_i \cap D'_j| = \begin{cases} |X_i \cap D_j| + N_{ij}, i = 1, 2, \cdots, m; j = 1, 2, \cdots, n \\ N_{ij}, i = m+1, \cdots, m+l \text{ 或 } j = n+1, \cdots, n+r \end{cases}$$

同时, $|X'_i| = \begin{cases} |X_i| + N_i, i = 1, 2, \cdots, m \\ N_i, i = m+1, \cdots, m+l \end{cases}$

特别地, 当进入系统为单一对象, 若其所在的条件类和决策类为 X_i 和 D_j 时, 由公式(6)可得:

$$E^{(t+1)}(D|C) = \frac{1}{(|U|+1)^2}(|U|^2 E^{(t)}(D|C) + 2|X_i - D_j|) \tag{10}$$

此时, 公式(10)结论同文献[3]。

当进入系统的对象集为相同条件类和决策类, 其所在的条件类和决策类为 X_i 和 D_j 时:

$$E^{(t+1)}(D|C) = \frac{1}{(|U|+N_i)^2}(|U|^2 E^{(t)}(D|C) + 2N_i|X_i - D_j|) \tag{11}$$

3.2 仅有对象集退出时条件熵的变化机制

假设 $t+1$ 时刻, 有 M 个对象集退出, $U'/C = \{X'_1, X'_2, \cdots, X'_m\}$, $U'/D = \{D'_1, D'_2, \cdots, D'_n\}$, 其中 U' 为退出对象集后系统的论域。$\forall X'_i \in U'/C$, 该条件类对应退出对象集数为 M_i, 而 $\forall D'_j \in U'/D$, 规则: $X'_i \to D'_j$ 中支持度减少的对象数为 M_{ij}, 示意图可详见文献[10, 11]。

基于以上的假设条件, 按照公式(5), 先分析 $t+1$ 时刻矩阵中的元素:

$$E(D'_j \mid X'_i) = |X'_i||X'_i \cap D'_j| - |X'_i \cap D'_j|^2$$
$$= (|X_i| - M_i)(|X_i \cap D_j| - M_{ij}) - (|X_i \cap D_j| - M_{ij})^2$$
$$= (|X_i||X_i \cap D_j| - |X_i \cap D_j|^2) - (M_{ij}|X_i| + M_i|X_i \cap D_j| - M_iM_{ij} - (2M_{ij}|X_i \cap D_j| - M_{ij}^2))$$

因此, $t+1$ 时刻条件熵 $E^{(t+1)}(D \mid C)$ 为:

$$E^{(t+1)}(D \mid C) = \frac{1}{(|U| - \sum_{i=1}^{m} M_i)^2}(|U|^2 E^{(t)}(D \mid C) + \sum_{i=1}^{m}((M_i^2 - 2M_i|X_i|) - \sum_{j=1}^{n}(M_{ij}^2 - 2M_{ij}|X_i \cap D_j|))) \tag{12}$$

其中: $M_i = \sum_{j=1}^{n} M_{ij}$, $M = \sum_{i=1}^{m} M_i$。此时, 条件熵求解的关键在于 M_{ij} 和 $|X_i \cap D_j|$, 根据公式(12)构造出增量矩阵 $Inc^{(t+1)}(D \mid C)$ 为:

$$Inc^{(t+1)}(D \mid C) = \begin{bmatrix} -M_{11} & -M_{12} & \cdots & -M_{1n} \\ -M_{21} & -M_{22} & \cdots & -M_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ -M_{m1} & -M_{m2} & \cdots & -M_{mn} \end{bmatrix} \tag{13}$$

原系统支持度矩阵为公式(7)。按照公式(12), 原有条件熵再结合公式(13)和(7)可得 $t+1$ 时刻条件熵。本文条件熵需计算: $mn(mc+n+1)$ 次。

$t+1$ 时刻, 支持度矩阵为 $Supp^{(t+1)}(D \mid C)$ 为:

$$Supp^{(t+1)}(D \mid C) = Supp^{(t)}(D \mid C) + Inc^{(t+1)}(D \mid C)$$
$$= \begin{bmatrix} |X_1 \cap D_1| - M_{11} & |X_1 \cap D_2| - M_{12} & \cdots & |X_1 \cap D_n| - M_{1n} \\ |X_2 \cap D_1| - M_{21} & |X_2 \cap D_2| - M_{22} & \cdots & |X_2 \cap D_n| - M_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ |X_m \cap D_1| - M_{m1} & |X_m \cap D_2| - M_{m2} & \cdots & |X_m \cap D_n| - M_{mn} \end{bmatrix} \tag{14}$$

由公式(14)可得, 矩阵中的元素为: $|X'_i \cap D'_j| = |X_i \cap D_j| - M_{ij}$。同时, $|X'_i| = |X_i| - M_i$。

特别地, 当单一对象退出系统且其所在的条件类和决策类分别为 X_i 和 D_j 时, 由公式(12)得:

$$E^{(t+1)}(D \mid C) = \frac{1}{(|U| - 1)^2}(|U|^2 E^{(t)}(D \mid C) - 2|X_i - D_j|) \tag{15}$$

3.3 对象集进入和退出共存时条件熵的变化机制

结合前面 3.1 和 3.2 部分讨论的情况, 这里假设在 $t+1$ 时刻, 系统中新增 N 个对象集, 同时有 M 个对象集退出, 采用前面同样地处理方法, 构造出增量矩阵 $Inc^{(t+1)}(D \mid C)$:

$$Inc^{(t+1)}(D \mid C) = \begin{bmatrix} N_{11} - M_{11} & N_{12} - M_{12} & \cdots & N_{1n} - M_{1n} & \cdots & N_{1(n+r)} \\ N_{21} - M_{21} & N_{22} - M_{22} & \cdots & N_{2n} - M_{2n} & \cdots & N_{2(n+r)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ N_{m1} - M_{m2} & N_{m2} - M_{m2} & \cdots & N_{mn} - M_{mn} & \cdots & N_{m(n+r)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ N_{(m+l)1} & N_{(m+l)2} & \cdots & N_{(m+l)n} & \cdots & N_{(m+l)(n+r)} \end{bmatrix} \tag{16}$$

在 $t+1$ 时刻, 条件熵 $E^{(t+1)}(D \mid C)$ 为:

$$E^{(t+1)}(D \mid C) = \frac{1}{(|U| + \sum_{i=1}^{m+l} N_i - M_i)^2} \cdot (|U|^2 E^{(t)}(D \mid C) + \sum_{i=1}^{m+l}(((N_i - M_i)^2 + 2(N_i - M_i)|X_i|) - \sum_{j=1}^{n+r}((N_{ij} - M_{ij})^2 + 2(N_{ij} - M_{ij})|X_i \cap D_j|))) \tag{17}$$

当系统中只有新增对象集或者只有退出对象集时, 公式(17)的结论显然同公式(6)和(12)。$t+1$ 时刻, 支持度矩阵 $Supp^{(t+1)}(D \mid C)$ 计算处理同前面 3.1 部分, 由于篇幅限制这里省略。按照公式(17),

原有条件熵再结合公式(16) 和(7) 可得 $t+1$ 时刻条件熵。

4　动态信息系统中核属性计算方法

结合定理 1 和动态信息系统中条件熵的变化机制, 下面考虑系统中同时存在对象集进入和退出的一般情况, 给出动态信息系统中一种基于条件熵的核属性计算方法。

算法: 动态信息系统中一种基于条件熵的核属性计算方法

输入: (1) t 时刻决策系统 $S=(U, C\cup D, V, f)$, 条件类和决策类为 $U/C=\{X_1, X_2, \cdots, X_m\}$, $U/D=\{D_1, D_2, \cdots, D_n\}$, 支持度矩阵 $Supp^{(t)}(D\mid C)$ 和条件熵 $E^{(t)}(D\mid C)$。对于每个 $a\in C$: 条件类 $U/(C-\{a\})=\{Z_1, Z_2, \cdots, Z_r\}(r\leqslant m)$, 而决策类保持不变, 支持度矩阵 $Supp^{(t)}(D\mid(C-\{a\}))$ 和条件熵 $E^{(t)}(D\mid(C-\{a\}))$。核属性 $CORE_D^{(t)}(C)$。

(2) $t+1$ 时刻, 新增的对象集 N 和退出系统的对象集 M。

输出: $t+1$ 的核属性 $CORE_D^{(t+1)}(C)$。

步骤 1: 对进入和退出系统的对象集, 按照公式(16) 构造出 $t+1$ 时刻的增量矩阵 $Inc^{(t+1)}(D\mid C)$。

步骤 2: 基于 $E^{(t)}(D\mid C)$, $Supp^{(t)}(D\mid C)$ 和 $Inc^{(t+1)}(D\mid C)$, 按公式(17) 计算出 $E^{(t+1)}(D\mid C)$。

步骤 3: 同理, 对于每一个 $a\in C$:

(1) 按照公式(16) 构造出 $t+1$ 时刻的增量矩阵 $Inc^{(t+1)}(D\mid(C-\{a\}))$, 再结合 $E^{(t)}(D\mid(C-\{a\}))$ 和支持度矩阵 $Supp^{(t)}(D\mid(C-\{a\}))$, 按照公式(17) 计算出 $E^{(t+1)}(D\mid(C-\{a\}))$。

(2) 若 $E^{(t+1)}(D\mid(C-\{a\}))-E^{(t+1)}(D\mid C)>0$, 则 $CORE=CORE\cup\{a\}$。

步骤 4: $t+1$ 的核属性 $CORE_D^t(C)=CORE$。

5　实例分析

下面将通过一个石油项目投资的风险评估系统来阐述本文所提出的算法。为便于同文献[3] 的算法进行比较, 这里只对新增对象集的情形进行实例分析。对于石油项目投资的风险指标, 主要有储量风险、地质和技术风险、财务风险、环境风险以及政策风险[14], 具体实例如表 1 所示。

表 1 中, 条件属性 C 有: 储量风险、地质和技术风险、财务风险、环境风险以及政策风险, 分别记为 a_1, a_2, a_3, a_4 和 a_5。因此, $C=\{a_1, a_2, a_3, a_4, a_5\}$。决策属性 D 为风险级别。表中数量列代表系统中对应论域的对象数。

表 **1**　关于石油项目投资的风险评估决策表

论域	储量风险	地质和技术风险	金融风险	环境风险	政策风险	风险级别	数量
1	1	1	2	1	1	1	5
2	1	2	2	2	1	1	5
3	1	2	2	2	1	2	10
4	2	2	3	2	2	2	2
5	3	2	3	2	2	2	15
6	3	2	3	2	2	1	2
7	3	2	3	2	2	2	3
8	4	2	4	2	3	2	5
9	4	2	4	2	3	3	5
10	4	2	4	2	3	3	10

在分析本文算法之前, 先分别计算出 t 时刻的等价类: $U/C=\{1, \{2, 3\}, 4, \{5, 6, 7\}, \{8, 9, 10\}\}$, $U/D=\{\{1, 2, 6\}, \{3, 4, 5, 7, 8\}, \{9, 10\}\}$。支持度矩阵 $Supp^{(t)}(D\mid C)$ 为:

$$Supp^{(t)}(D\mid C)=\begin{bmatrix}5&0&0\\5&10&0\\0&2&0\\2&18&0\\0&5&15\end{bmatrix}。$$

按照公式(4), 求得 $E^{(t)}(D\mid C)=161/1922$。同理对于每个 $a\in C$, 可以构造出支持度矩阵 $Supp^{(t)}(D\mid(C-\{a\}))$ 和条件熵 $E^{(t)}(D\mid(C-\{a\}))$。结果如下:

$$Supp^{(t)}(D\mid(C-\{a_1\}))=\begin{bmatrix}5&0&0\\5&10&0\\2&20&0\\0&5&15\end{bmatrix}。$$

因此, $E^{(t)}(D\mid(C-\{a_1\}))=165/1922$。

$Supp^{(t)}(D\mid(C-\{a_2\}))$, $Supp^{(t)}(D\mid(C-\{a_3\}))$, $Supp^{(t)}(D\mid(C-\{a_4\}))$ 和 $Supp^{(t)}(D\mid(C-\{a_5\}))$ 都与 $Supp^{(t)}(D\mid C)$ 相同。

$E^{(t)}(D\mid(C-\{a_2\}))=161/1922$,

$E^{(t)}(D\mid(C-\{a_3\}))=161/1922$,

$E^{(t)}(D\mid(C-\{a_4\}))=161/1922$,

$E^{(t)}(D\mid(C-\{a_5\}))=161/1922$。

此时, 核属性为: $CORE_D^{(t)}(C)=\{a_1\}$。

假设 $t+1$ 时刻, 系统出现新增对象集, 如表 2 所示。

由新增对象集可知, 系统中新增一个条件类和一个决策。$U'/C=\{\{1, 13\}, \{2, 3\}, 4, \{5, 6, 7, 12\}, \{8, 9, 10\}, 11\}$, $U'/D=\{\{1, 2, 6, 13\}, \{3, 4, 5, 7, 8, 12\}, \{9, 10\}, 11\}$, 其中 U' 为新增对象集后系统

的论域

表 **2** 新增对象集

论域	储量风险	地质和技术风险	金融风险	环境风险	政策风险	风险级别	数量
11	3	4	3	2	4	4	4
12	3	2	3	2	2	2	5
13	1	1	2	1	1	1	5

按照本文算法中的步骤 1, 构造出增量矩阵:

$$Inc^{(t+1)}(D \mid C) = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}。$$

步骤 2: 由 $Inc^{(t+1)}(D \mid C)$ 中新增对象集在各条件类和决策类上的分布, 找出 $Supp^{(t)}(D \mid C)$ 中对应的交集, 加上 $E^{(t)}(D \mid C)$ 项, 可得条件熵: $E^{(t+1)}(D \mid C) = 171/2888$。下面详细列出此时条件熵在本文的算法和文献[3] 的算法中的具体计算过程, 如表 3 所示。

表 **3** 条件熵在两种算法下的具体计算过程

算法	具体计算过程
本文的算法	(1) 按增量矩阵的非零元素所在行的顺序, 结合支持度矩阵, 依次可得: 第一行: $5^2 + 2 \times 5 \times 5 - (5^2 + 2 \times 5 \times 5) = 0$ 第四行: $5^2 + 2 \times 5 \times 20 - (5^2 + 2 \times 5 \times 18) = 20$ 第六行: $4^2 + 2 \times 4 \times 0 - (4^2 + 2 \times 4 \times 0) = 0$ (2) 加入 t 时刻的条件熵, 由此可得新的条件熵为: $E^{(t+1)}(D \mid C) = \frac{1}{(62+14)^2}(62^2 \times \frac{161}{1922} + 0 + 20 + 0)$ 因此, 可得 $E^{(t+1)}(D \mid C) = 171/2888$。
文献[3]的算法	论域中 11 所对对应第 1 个对象, 进入系统, 再比较系统中条件类与决策类后可得: $E^{(t+1)}(D \mid C) = \frac{1}{(62+1)^2}(62^2 \times \frac{161}{1922} + 2 \times 0) = \frac{322}{63^2}$ 第 2 个对象进入系统, 再比较系统中的条件类与决策类后可得: $E^{(t+1)}(D \mid C) = \frac{1}{(63+1)^2}(63^2 \times \frac{312}{63^2} + 2 \times 0) = \frac{322}{64^2}$ 第 3 个对象进入系统, 再比较系统中的条件类与决策类后可得: $E^{(t+1)}(D \mid C) = \frac{1}{(64+1)^2}(64^2 \times \frac{312}{64^2} + 2 \times 0) = \frac{322}{65^2}$ 同理, 论域中其他对象集按单个对象依次进入系统且比较系统中的条件类与决策类后, 最终可得条件熵: $E^{(t+1)}(D \mid C) = \frac{1}{(75+1)^2}(75^2 \times \frac{342}{75^2} + 2 \times 0) = \frac{171}{2888}$

表 3 中显示出本文的算法同文献[3] 中的算法在处理对象集方面明显的差异, 前者是采用批量处理, 而后者是单个对象依次处理。

这里以条件熵 $E^{(t+1)}(D \mid C)$ 的求解为例, 通过计算次数来对比分析本文算法和文献[3] 中的算法:

本文的算法: 需要 $(5 \times 5 + 3 \times 1) \times 3 + 4 \times 6 = 108$ 次。

文献[3] 的算法: 需要 $(5 \times 5 + 3 \times 1) + (6 \times 5 + 4 \times 1) \times (3 + 5 + 5) = 470$ 次。

此时, 本文的算法相对文献[3] 的算法在求解条件熵所需的比较次数要少, 具有一定优势。造成文献[3] 的算法比较次数多的原因, 有两点: (1) 存在新的条件类或决策类先进入系统; (2) 该方法计算时是按单个对象依次进行。

依次, 按照本文算法的步骤 3 可以分别求得 $t+1$ 时刻, $E^{(t+1)}(D \mid (C - \{a_1\})) = 175/1922$,

$E^{(t+1)}(D \mid (C - \{a_2\})) = 171/2888$,

$E^{(t+1)}(D \mid (C - \{a_3\})) = 171/2888$,

$E^{(t+1)}(D \mid (C - \{a_4\})) = 171/2888$,

$E^{(t+1)}(D \mid (C - \{a_5\})) = 171/2888$。

此时, 核属性为: $CORE_b^{(t+1)}(C) = \{a_1\}$。

为便于动态系统下一次求解核属性, 需重新构造出 $t+1$ 时刻, 各属性组合所对应的支持度矩阵。这里以全部属性都存在下的支持度矩阵为例:

$$Supp^{(t+1)}(D \mid C) = Supp^{(t)}(D \mid C) + Inc^{(t+1)}(D \mid C)$$

$$= \begin{bmatrix} 5 & 0 & 0 & 0 \\ 5 & 10 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 2 & 18 & 0 & 0 \\ 0 & 5 & 15 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 0 & 0 & 0 \\ 5 & 10 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 2 & 23 & 0 & 0 \\ 0 & 5 & 15 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}。$$

对于动态系统存在对象集退出或者对象集进入和退出同时存在的情形, 处理方式类似上面的实例分析, 由于篇幅限制这里省略。

6 结语

针对实际动态信息系统存在新增和退出对象集的情况, 文中分别讨论仅有对象集进入、仅有对象集

退出以及对象集进入和退出同时存在下条件熵的变化机制。研究发现: $t+1$ 时刻条件熵求解关键在于新增或退出对象集在各条件类和决策类下的分布和原有条件类与决策类的交集。基于此,构造支持度矩阵和增量矩阵进而提出了动态系统中一种基于条件熵的核属性计算方法。本文方法适应于批量处理新增和退出对象集,若新增对象集有新的条件类或者决策类情况该方法可能也具有一定优势。

参考文献:

[1] Blaszczynski, J., Slowinski, R.. Incremental induction of decision rules from dominance-based rough approximation[J]. Electronic Notes in Theoretical Computer Science, 2003, 82:40-51.

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

[3] 梁吉业, 魏巍, 钱宇华. 一种基于条件熵的增量核求解方法[J]. 系统工程理论与实践, 2008, 4: 81-89.

[4] 骆公志, 杨晓江. 变精度优势粗糙集属性约简择优算法[J]. 中国管理科学, 2009, 17(2): 169-175.

[5] Hu, X. Y., Gercone, N.. Learning in relational database: A rough set approach[J]. Computational Intelligence, 1995, 11 (2): 339-347.

[6] 叶英毅, 陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7): 1086-1088.

[7] 杨明, 孙志挥. 改进的差别矩阵及其求核方法[J]. 复旦大学学报(自然版), 2004, 43(5): 865-868.

[8] 王国胤. 决策表核属性的计算方法[J]. 计算机学报, 2003, 26(5): 611-615.

[9] Wang, G. Y., Zhao, J., An, J. J., et al. A comparative study of algebra viewpoint and information viewpoint in attribute reduction[J]. Fundamenta Informaticae, 2005, 68: 289-301.

[10] Liu, D., Li, T. R., Ruan, D., et al. An incremental approach for inducing knowledge from dynamic information systems[J]. Fundamenta Informaticae, 2009, 94: 245-260.

[11] Liu, D., Li, T. R., Ruan, D., et al. Incremental learning optimization on knowledge discovery in dynamic business intelligent systems [J]. Journal of Global Optimization, DOI: 10. 1007/s10898-010-9607-8.

[12] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J].计算机学报, 2006, 29(3): 407-413.

[13] 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.

[14] Xie, G., Yue, W. Y., Wang, S. Y., et al. Dynamic risk management in petroleum project investment based on a variable precision rough set model [J]. Technological Forecasting and Social Change, 2010, 77: 891-901.

A Calculation Method for Core Attributes Based on Conditional Entropy in Dynamic Information Systems

LIANG De-cui, HU Pei

(School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: The situation that some objects immigrate the system and some objects emigrate the system simultaneously exists in dynamical information systems. In this paper, a calculation method for core attributes is proposed in information view. First, the changing mechanism of condition entropy is analyzed from three different cases, whick include the objects' immigration or emigration and that this two cases coexisted. Based on the mechanism, the new condition entropy is computed quickly by support matrixes and incremental matrixes. These matrixes can constructed by computing the objects immigrated or emmigrated at the same time. Then core attributes are obtained quickly. Finally, the validity of the method has been depicted by a practical example.

Key words: dynamic information systems; incremental matrix; support matrix; conditional entropy; core attributes