文章编号: 1000-6788(2008)04-0081-09

一种基于条件熵的增量核求解方法

梁吉业[a,b], 魏　巍[a,b], 钱宇华[a,b]
(山西大学 a. 计算机与信息技术学院, b. 智能信息处理研究所, 太原 030006)

摘要:　分析了增加新对象后, 决策表的决策属性关于条件属性的条件熵变化原理. 并在此基础上提出了一种新的增量核求解算法. 该算法只需找到与新对象属性值相等的条件类和决策类就可以得到新的条件熵, 进而求得决策表在信息观下的增量属性核. 实例证明了该算法的有效性.

关键词:　决策表; 条件熵; 核; 增量算法
中图分类号:　TP18　　文献标志码:　A

An incremental approach to computation of a core based on conditional entropy

LIANG Jiye[a,b], WEI Wei[a,b], QIAN Yuhua[a,b]
(a. School of Computer & Information Technology, b. Research Institute of Intelligent Information Processing, Shanxi University, Taiyuan 030006, China)

Abstract: In this paper, the changing mechanism of conditional entropy is analyzed when a new object is added to the original decision table. Based on the mechanism, a new incremental algorithm of the computation of a core is proposed. By means of this algorithm, the computation of new conditional entropy need only to find the condition class and the decision class with the equal attribute value of the element newly added to a decision table. Furthermore, incremental attribute core in information view can be calculated in a decision table. Finally, the validity of the algorithm have been depicted by an practical example.

Key words: decision table; conditional entropy; core; incremental algorithm

1　引言

粗糙集理论作为一种处理不精确、不确定和不完全数据的数学工具, 已被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. 粗糙集得到的知识是人们可以理解的关联规则, 这些关联规则符合人类的经验, 更适合在管理决策中应用. 因此, 20 多年来, 粗糙集理论得到了迅速的发展.

决策表的属性约简是粗糙集理论中的核心概念之一[1~3], 属性核的求解又是多数启发式属性约简算法的关键步骤. 因此, 研究求解属性核的有效方法具有重要的理论意义和应用价值. 目前已有许多求解属性核的方法[4~11]. 文献[6~9]对属性核求解问题进行了深入的探讨, 证明了在协调的决策表中粗糙集理论代数观(以下简称为"代数观")下的属性核与粗糙集理论信息观(以下简称为"信息观")下的属性核是一致的, 在不协调的决策表中信息观下的属性核包含代数观下的属性核. 文献[10]提出的差别矩阵方法是经典的属性核求解方法之一. 该方法可有效地减少计算量, 提高求解属性核的效率, 但在不协调的决策表中得不到正确的属性核. 于是, 文献[11]在文献[10]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, 该方法可以得到不协调决策表代数观下的属性核. 文献[12]提出了一种新的改进的差别

收稿日期: 2006-12-01
资助项目: 国家自然科学基金(70471003, 60773133); 国家 863 计划项目(2007AA01Z165); 教育部高等学校博士点专项基金(20050108004); 教育部科学技术研究重点项目(206017); 山西省重点实验室开放基金(200603023)

作者简介: 梁吉业(1962-), 男, 教授, 博士生导师, 研究方向: 计算智能, 智能决策; 魏巍(1980-), 男, 博士研究生, 研究方向: 智能决策, 粗糙集理论; 钱宇华(1976-), 男, 博士研究生, 研究方向: 粒度计算, 智能决策.

矩阵及其求核方法, 该方法可以得到代数观下的属性核, 而且能有效地降低计算复杂度. 文献[6] 提出了一种基于信息熵的属性核求解算法, 该算法可以得到信息观下属性核, 有效地补充和完善了文献[11] 中的算法, 但算法效率仍需改进. 可见, 前人已在属性核求解的问题上做了大量的工作, 但是关于求解增量属性核的方法却报道不多.

然而现实世界中, 数据库的规模不断增大, 数据库的更新导致了其信息结构的变化, 已有的属性核可能不再有效, 这就需要对其进行动态修改. 文献[13] 提出一种基于改进差别矩阵的核属性增量式更新算法, 该算法在更新差别矩阵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, 该算法可以得到代数观下决策表的增量属性核. 本文利用文献[14, 15] 中提出的条件熵建立了决策表的对象增加与决策属性关于条件属性的条件熵变化之间的定量关系, 提出了一种基于条件熵的增量核求解算法, 该算法可以得到信息观下的增量属性核, 实例证明了该算法的有效性.

2 代数观和信息观下的属性核

2.1 代数观下的属性核

定义 **1** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, C 在 U 上产生的划分为 $U/C=\{X_1,X_2,\cdots,X_m\}$, D 在 U 上产生的划分为 $U/D=\{Y_1,Y_2,\cdots,Y_n\}$, D 的 C 正域定义为

$$POS_C(D)=\bigcup_{i=1}^{n}\underline{C}Y_i.$$

定义 **2** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, 对于 $\forall a\in C$, 若 $POS_C(D)\neq POS_{C-\{a\}}(D)$, 则属性 a 属于代数观下 D 的 C 核 $CORE_D^A(C)$.

2.2 信息观下的属性核

定义 **3**[14] 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, $U/C=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{Y_1,Y_2,\cdots,Y_n\}$, 决策表的条件熵定义为

$$E_U(D\mid C)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Y_i\cap X_j|}{|U|}\,\frac{|Y_i^c-X_j^c|}{|U|}.$$

其中, $E_U(D\mid C)$ 的下标 U 表示论域, X_j^c 表示 X_j 的补集, Y_i^c 表示 Y_i 的补集.

定理 **1**[14] 设 $S_1=(U,C,V,f)$ 和 $S_2=(U,C',V,f)$ 是两个信息系统, D 是 U 上的一个决策. 如果 $C'\subseteq C$, 则 $E(D\mid C)\leqslant E(D\mid C')$.

定理 **2** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, 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 $CORE_D^I(C)$ 的充分必要条件为

$$E_U(D\mid C)<E_U(D\mid C-\{a\}).$$

证明 充分性: 如果 $E_U(D\mid C)<E_U(D\mid C-\{a\})$, 则根据定理 1 可知属性集 $C-\{a\}$ 的任意子集都不可能是 C 的约简, 所以 a 包含于所有的属性约简中, 即 a 为信息观下的核属性.

必要性: 如果 a 为核属性, 则 $C-\{a\}\subset C$, 根据定理 1 可得 $E_U(D\mid C)\leqslant E_U(D\mid C-\{a\})$. 如果 $E_U(D\mid C)=E_U(D\mid C-\{a\})$, 根据约简的定义一定有 $A\subseteq C-\{a\}$ 是 C 的约简, 又因为 a 为核属性, 所以 $\{a\}\in A$, 这与 $A\subseteq C-\{a\}$ 矛盾. 因此, $E_U(D\mid C)<E_U(D\mid C-\{a\})$.

证毕.

3 基于差别矩阵的增量属性核求解方法

3.1 基于差别矩阵的属性核求解方法

Hu X H 教授在文献[10] 中根据 Skowron A 教授在文献[4] 中提出的差别矩阵得出一种确定决策表属性核的方法. 叶东毅教授在文献[11] 中对文献[10] 的结论提出了质疑, 举例证明了其结论的问题, 并提出了一种基于改进差别矩阵的属性核求解方法, 但是, 他还没有发现产生这个问题的根本原因. 王国胤教授在文献[6~ 9] 中系统地研究和阐述了信息观下的属性核和代数观下的属性核的差异, 并提出了基于信息

熵的决策表属性核计算方法. 杨明教授在文献[12] 中对文献[11] 中的差别矩阵进行了改进, 即下面的定义 4.

定义 **4**[12]　对给定的决策表, 差别矩阵 $M_1 = \{m_{ij}\}$ 定义为

$$m_{ij} = \begin{cases} \{a \in C: f(x_i, a) \neq f(x_j, a)\}, & \text{当} f(x_i, D) \neq f(x_j, D), \text{且} x_i \in U_1, x_j \in U_1 \text{时} \\ \{a \in C: f(x_i, a) \neq f(x_j, a)\}, & \text{当} x_i \in U_1, x_j \in U_2 \text{时} \\ \varnothing & \text{其它} \end{cases}$$

其中 $U_1 \in \bigcup_{i=1}^{n} \underline{C}Y_i$, $U_2 = U - U_1$.

杨明教授提出的差别矩阵与叶东毅教授的差别矩阵实质上都是只考虑了 $x_i \in U_1$, $x_j \in U_1$ 和 $x_i \in U_1$, $x_j \in U_2$ 的情况, 而没有考虑 $x_i \in U_2$, $x_j \in U_2$. 因此杨明教授提出的基于其差别矩阵的算法对于不协调的决策表只能得到代数观下的属性核.

3.2　基于差别矩阵的增量属性核求解方法

杨明教授在文献[13] 中以文献[12] 的定义 4 为基础, 进一步提出定义 5 和定理 3.

定义 **5**[13]　对于给定的决策表, 差别矩阵 $M_2 = \{m_{ij}\}$ 定义为

$$m_{ij} = \begin{cases} \{a \in C: f(x_i, a) \neq f(x_j, a)\} & \text{当} f(x_i, D) \neq f(x_j, D), \text{且} x_i \in U_1, x_j \in U_1 \text{时} \\ \{a \in C: f(x_i, a) \neq f(x_j, a)\} & \text{当} x_i \in U_1, x_j \in U'_2 \text{时} \\ \varnothing & \text{其他} \end{cases}$$

其中 $U_1 \in \bigcup_{i=1}^{n} \underline{C}Y_i$, $U_2 = U - U_1$, U'_2 表示从 U_2 的每个条件类中任取一个对象组成的新对象集.

定理 **3**[13]　对于给定的决策表, 若记 $IDM(C, M_2) = \{m_{ij} \mid m_{ij} \in M_2 \text{ 且 } m_{ij} \text{ 为单个属性}\}$, 则有 $IDM(C, M_2) = CORE_D^A(C)$, 即当且仅当某个 m_{ij} 为单个属性时, 该属性属于核 $CORE_D^A(C)$.

杨明教授根据定义 5 和定理 3 提出了一种基于改进差别矩阵的核增量式更新算法. 其基本思想如下:

设决策表动态改变时, 决策属性 D 的取值范围仍为 1, 2, …, k. 对新增对象 x, 若 $\forall y \in (U_1 \cup U'_2)$, x 和 y 协调的, 则称 x 与 $(U_1 \cup U'_2)$ 协调; 若 $\exists y \in U_1$ 使得 x 和 y 是不协调的, 则称 x 与 U_1 不协调; 若 $\exists y \in U'_2$ 使得 x 和 y 是不协调的, 则称 x 与 U'_2 不协调.

由定义 5 得到差别矩阵为 M_2, 若新增对象为 x, 则只要得到 $(U_1 \cup U'_2 \cup \{x\})$ 的差别矩阵, 便可由定理 3 求得属性核. 因此, 增量属性核的求解本质上就是差别矩阵的更新问题. M_2 的更新可分为下列 3 种情况:

1) 若 x 与 $(U_1 \cup U'_2)$ 协调, 则在 M_2 中增加对象 x 对应的行和列, $U_1 = U_1 \cup \{x\}$;

2) 若 x 与 U_1 不协调, 则在 M_2 中删除对象 y 所在的行, 修改对象 y 所在的列, $U'_2 = U'_2 \cup \{y\}$, $U_1 = U_1 - \{y\}$;

3) 若 x 与 U'_2 不协调, 则 M_2 保持不变.

文献[13] 提出的增量核求解算法是以文献[12] 中的差别矩阵为基础的, 因而该算法得到的是代数观下的增量属性核.

4　新增对象后条件熵的变化机制

给定决策表 $S = (U, R, V, f)$, 其中 $R = C \cup D$, C 是条件属性集, D 是决策属性集, $A \subseteq C$. 新对象 x 加入决策表后, 决策属性关于条件属性的条件熵的变化可以分为以下四种情况分析.

1) x 不属于 U/A 中的条件类, x 不属于 U/D 中的决策类;

2) x 不属于 U/A 中的条件类, x 属于 U/D 中的决策类;

3) x 属于 U/A 中的条件类, x 不属于 U/D 中的决策类;

4) x 属于 U/A 中的条件类, x 属于 U/D 中的决策类.

这四种情况都可以利用定理 4 中的统一公式得到新的条件熵.

定理 **4** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, $A\subseteq C$. $U/A=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{Y_1,Y_2,\cdots,Y_n\}$, 增加一个对象 x 后, $x\in X'_q$ 且 $x\in Y'_p$ (X'_q 是 $(U\cup\{x\})/A$ 中的条件类, Y'_p 是 $(U\cup\{x\})/D$ 中的决策类), 则新的条件熵为

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(|U|+1)^2}(|U|^2E_U(D\mid A)+2|X'_q-Y'_p|).$$

为了证明定理 4, 我们首先证明引理 1~ 4.

引理 **1** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, $A\subseteq C$. $U/A=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{Y_1,Y_2,\cdots,Y_n\}$. 增加一个对象 x 后, $(U\cup\{x\})/A=\{X'_1,X'_2,\cdots,X'_m,X'_{m+1}\}$, $(U\cup\{x\})/D=\{Y'_1,Y'_2,\cdots,Y'_n,Y'_{n+1}\}$, 若 $x\in X'_{m+1}$ ($X'_{m+1}=\{x\}$), 且 $x\in Y'_{n+1}$ ($Y'_{n+1}=\{x\}$), 则

$$E_{U\cup\{x\}}(D\mid A)=\frac{|U|^2}{(|U|+1)^2}E_U(D\mid A).$$

证明略.

引理 **2** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, $A\subseteq C$. $U/A=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{Y_1,Y_2,\cdots,Y_n\}$. 增加一个对象 x 后, $(U\cup\{x\})/A=\{X'_1,X'_2,\cdots,X'_m,X'_{m+1}\}$, $(U\cup\{x\})/D=\{Y'_1,Y'_2,\cdots,Y'_n\}$, 若 $x\in X'_{m+1}$ ($X'_{m+1}=\{x\}$), 且 $x\in Y'_p$ ($Y'_p=Y_p\cup\{x\}$, $p\leqslant n$), 则

$$E_{U\cup\{x\}}(D\mid A)=\frac{|U|^2}{(|U|+1)^2}E_U(D\mid A).$$

证明略.

引理 **3** 给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, $A\subseteq C$. $U/A=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{Y_1,Y_2,\cdots,Y_n\}$. 增加一个对象 x 后, $(U\cup\{x\})/A=\{X'_1,X'_2,\cdots,X'_m\}$, $(U\cup\{x\})/D=\{Y'_1,Y'_2,\cdots,Y'_n,Y'_{n+1}\}$, 若 $x\in X'_q$ ($X'_q=X_q\cup\{x\}$, $q\leqslant m$), 且 $x\in Y'_{n+1}$ ($Y'_{n+1}=\{x\}$), 则

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(|U|+1)^2}(|U|^2E_U(D\mid A)+2|X_q|).$$

证明　根据已知条件可得 $X'_1=X_1,\cdots,X'_q=X_q\cup\{x\},\cdots,X'_m=X_m$, $Y'_1=Y_1,\cdots,Y'_n=Y_n$, $Y'_{n+1}=Y'_p=\{x\}$. 则

$$\begin{aligned}
E_{U\cup\{x\}}(D\mid A)&=\sum_{i=1}^{n+1}\sum_{j=1}^{m}\frac{|Y'_i\cap X'_j|}{|U\cup\{x\}|}\frac{|Y'^c_i-X'^c_j|}{|U\cup\{x\}|}\\
&=\sum_{i=1}^{n}\sum_{j=1,j\neq q}^{m}\frac{|Y'_i\cap X'_j|}{|U\cup\{x\}|}\frac{|Y'^c_i-X'^c_j|}{|U\cup\{x\}|}+\sum_{i=1}^{n}\frac{|Y'_i\cap X'_q|}{|U\cup\{x\}|}\frac{|Y'^c_i-X'^c_q|}{|U\cup\{x\}|}\\
&\quad+\sum_{j=1}^{m}\frac{|Y'_{n+1}\cap X'_j|}{|U\cup\{x\}|}\frac{|Y'^c_{n+1}-X'^c_j|}{|U\cup\{x\}|}\\
&=\sum_{i=1}^{n}\sum_{j=1,j\neq q}^{m}\frac{|Y_i\cap X_j||Y_i^c-X_j^c|}{(|U|+1)^2}+\sum_{i=1}^{n}\frac{|Y_i\cap(X_q\cup\{x\})||Y_i^c-(X_q\cup\{x\})^c|}{(|U|+1)^2}\\
&\quad+\sum_{j=1,j\neq q}^{m}\frac{|\{x\}\cap X_i||\{x\}^c-X_i^c|}{(|U|+1)^2}+\frac{|\{x\}\cap(X_q\cup\{x\})||\{x\}^c-(X_q\cup\{x\})^c|}{(|U|+1)^2}\\
&=\sum_{i=1}^{n}\sum_{j=1,j\neq q}^{m}\frac{|Y_i\cap X_j||Y_i^c-X_j^c|}{(|U|+1)^2}+\sum_{i=1}^{n}\frac{|Y_i\cap X_q||Y_i^c-X_q^c|+|Y_i\cap X_q||\{x\}|}{(|U|+1)^2}\\
&\quad+\frac{|X_q||\{x\}|}{(|U|+1)^2}\\
&=\frac{1}{(|U|+1)^2}(|U|^2E_U(D\mid A)+2|X_q|)
\end{aligned}$$

证毕.

引理 **4**　给定决策表 $S=(U,R,V,f)$, 其中 $R=C\cup D$, C 是条件属性集, D 是决策属性集, $A\subseteq C$. $U/A=\{X_1,X_2,\cdots,X_m\}$, $U/D=\{Y_1,Y_2,\cdots,Y_n\}$. 增加一个对象 x 后, $(U\cup\{x\})/A=\{X'_1,X'_2,\cdots,X'_m\}$, $(U\cup\{x\})/D=\{Y'_1,Y'_2,\cdots,Y'_n\}$, 若 $x\in X'_q(X'_q=X_q\cup\{x\},q\leqslant m)$, 且 $x\in Y'_p(Y'_p=Y_p\cup\{x\},p\leqslant n)$, 则

$$U_{U\cup\{x\}}(D\mid A)=\frac{1}{(|U|+1)^2}(|U|^2E_U(D\mid A)+2|X_q-Y_p|).$$

证明　根据已知条件可得 $X'_1=X_1,\cdots,X'_q=X_q\cup\{x\},\cdots,X'_m=X_m,Y'_1=Y_1,\cdots,Y'_p=Y_p\cup\{x\},\cdots,Y'_n=Y_n$. 则

$$\begin{aligned}
E_{U\cup\{x\}}(D\mid A)&=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Y'_i\cap X'_j|}{|U\cup\{x\}|}\frac{|Y'^c_i-X'^c_j|}{|U\cup\{x\}|}\\
&=\sum_{i=1,i\neq p}^{n}\sum_{j=1,j\neq q}^{m}\frac{|Y_i\cap X_j|}{|U\cup\{x\}|}\frac{|Y^c_i-X^c_j|}{|U\cup\{x\}|}+\sum_{i=1,i\neq p}^{n}\frac{|Y_i\cap X'_q|}{|U\cup\{x\}|}\frac{|Y^c_i-X'^c_q|}{|U\cup\{x\}|}\\
&\quad+\sum_{j=1}^{m}\frac{|Y'_p\cap X'_j|}{|U\cup\{x\}|}\frac{|Y'^c_p-X'^c_j|}{|U\cup\{x\}|}\\
&=\sum_{i=1,j\neq p}^{n}\sum_{1,j\neq q}^{m}\frac{|Y_i\cap X_j||Y^c_i-X^c_j|}{(|U|+1)^2}+\sum_{j=1,j\neq p}^{m}\frac{|(Y_p\cup\{x\})\cap X_j||Y_p\cup\{x\})^c-X^c_j|}{(|U|+1)^2}\\
&\quad+\sum_{i=1,i\neq p}^{n}\frac{|Y_i\cap(X_q\cup\{x\})||Y^c_i-(X_q\cup\{x\})^c|}{(|U|+1)^2}\\
&\quad+\frac{|(Y_p\cup\{x\})\cap(X_q\cup\{x\})||(Y_p\cup\{x\})^c-(X_q\cup\{x\})^c|}{(|U|+1)^2}\\
&=\sum_{i=1,i\neq p}^{n}\sum_{1,j\neq q}^{m}\frac{|Y_i\cap X_j||Y^c_i-X^c_j|}{(|U|+1)^2}+\sum_{j=1,j\neq p}^{m}\frac{|Y_p\cap X_j||Y^c_p-X^c_j|}{(|U|+1)^2}\\
&\quad+\sum_{i=1,i\neq p}^{n}\frac{|Y_i\cap X_q||Y^c_i-X^c_q|+|Y_i\cap X_q||\{x\}|}{(|U|+1)^2}\\
&\quad+\frac{|Y_p\cap X_q||Y^c_p-X^c_q|+|\{x\}||Y^c_p-X^c_q|}{(|U|+1)^2}\\
&=\frac{|U|^2}{(|U|+1)^2}E_U(D\mid A)+\frac{|(U-Y_p)\cap X_q|+|Y^c_p-X^c_q|}{(|U|+1)^2}\\
&=\frac{1}{(|U|+1)^2}(|U|^2E_U(D\mid A)+2|X_q-Y_p|)
\end{aligned}$$

证毕.

下面证明定理 4.

证明　1) 当 x 不属于 U/A 中的条件类, x 不属于 U/D 中的决策类时, 即 $x\in X'_{m+1}(X'_{m+1}=\{x\})$ 且 $x\in Y'_{n+1}(Y'_{n+1}=\{x\})$ 时.

根据引理 1 可得

$$E_{U\cup\{x\}}(D\mid A)=\frac{|U|^2}{(|U|+1)^2}E_U(D\mid A),$$

又因为 $X'_q=\{x\}$, $Y'_p=\{x\}$, 则,

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(|U|+1)^2}(|U|^2E_U(D\mid A)+2|X'_q-Y'_p|).$$

2) 当 x 不属于 U/A 中的条件类, x 属于 U/D 中的决策类时, 即 $x\in X'_{m+1}(X'_{m+1}=\{x\})$ 且 $x\in Y'_p(Y'_p=Y_p\cup\{x\},p\leqslant n)$ 时.

根据引理 2 可得

$$E_{U\cup\{x\}}(D\mid A)=\frac{|U|^2}{(|U|+1)^2}E_U(D\mid A),$$

又因为 $X'_q=\{x\}$, $Y'_p=Y_p\cup\{x\}$, 则

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid A)+2\mid X'_q-Y'_p\mid).$$

3) 当 x 属于 U/A 中的条件类, x 不属于 U/D 中的决策类时, 即 $x\in X'_q(X'_q=X_q\cup\{x\}, q\leqslant m)$ 且 $x\in Y'_{n+1}(Y'_{n+1}=\{x\})$ 时.

根据引理 3 可得

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid A)+2\mid X_q\mid),$$

又因为 $X'_q=X_q\cup\{x\}$, $Y'_p=\{x\}$, 则

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid A)+2\mid X'_q-Y'_p\mid).$$

4) 当 x 属于 U/A 中的条件类, x 属于 U/D 中的决策类时, 即 $x\in X'_q(X'_q=X_q\cup\{x\}, q\leqslant m)$ 且 $x\in Y'_p$ $(Y'_p=Y_p\cup\{x\}, p\leqslant n)$ 时,

根据引理 4 可得

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid A)+2\mid X_q-Y_p\mid),$$

又因为 $X'_q=X_q\cup\{x\}$, $Y'_p=Y_p\cup\{x\}$, 则

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid A)+2\mid X'_q-Y'_p\mid).$$

所以, 对于新对象 x 加入决策表的四种不同情况都满足

$$E_{U\cup\{x\}}(D\mid A)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid A)+2\mid X'_q-Y'_p\mid).$$

证毕.

5 决策表增量核求解算法

5.1 算法

结合定理 2 与定理 4 可以设计出一种新的增量核求解算法.

算法 基于条件熵的增量核求解算法(ICBCE)

输入: 1) 一个决策表 $S=(U, C\cup D, V, f)$, 现有决策表的核 $CORE^I_D(C)$;

2) 新增加的对象 x.

输出: 决策表新的核 $CORE^I_D(C)$.

步骤 **1** 在 $U/D=\{Y_1, Y_2, \cdots, Y_n\}$ 中查找与 x 决策属性值相等的决策类 Y_p, 若存在这样的决策类, 则 $Y_p:=Y_p\cup\{x\}(p\leqslant n)$, $U\cup\{x\}/D:=\{Y_1, Y_2, \cdots, Y_n\}$, 否则 $Y_p:=Y_{n+1}:=\{x\}$, $U\cup\{x\}/D:=\{Y_1, Y_2, \cdots, Y_n, Y_{n+1}\}$;

步骤 **2** 在 $U/C=\{X_1, X_2, \cdots, X_m\}$ 中查找与 x 各条件属性值都相等的条件类 X_q, 若存在这样的条件类, 则 $X_q:=X_q\cup\{x\}(q\leqslant m)$, $U\cup\{x\}/C:=\{X_1, X_2, \cdots, X_m\}$, 否则 $X_q:=X_{m+1}:=\{x\}$, $U\cup\{x\}/C:=\{X_1, X_2, \cdots, X_m, X_{m+1}\}$;

步骤 **3** 计算 $E_{U\cup\{x\}}(D\mid C)=\frac{1}{(\mid U\mid+1)^2}(\mid U\mid^2E_U(D\mid C)+2\mid X_q-Y_p\mid)$;

步骤 **4** 对于每一个 $a\in C$;

1) 在 $U/(C-\{a\})=\{Z_1, Z_2, \cdots, Z_r\}(r\leqslant m)$ 中查找各条件属性值与 x 都相等的等价类 Z_s, 若存在这样的等价类, 则 $Z_s:=Z_s\cup\{x\}(s\leqslant r)$, $U\cup\{x\}/(C-\{a\}):=\{Z_1, Z_2, \cdots, Z_r\}$, 否则 $Z_s:=Z_{r+1}:=\{x\}$, $U\cup\{x\}/(C-\{a\}):=\{Z_1, Z_2, \cdots, Z_r, Z_{r+1}\}$;

2) 计算 $E_{U\cup\{x\}}(D|C-\{a\})=\frac{1}{(|U|+1)^2}(|U|^2E_U(D|C-\{a\})+2|Z_s-Y_p|)$;

3) 若 $E_{U\cup\{x\}}(D|C-\{a\})-E_{U\cup\{x\}}(D|C)>0$, 则 $CORD_D^I(C):=CORD_D^I(C)\cup\{a\}$;

步骤 **5**　$U:=U\cup\{x\}$, 输出新核 $CORE_D^I(C)$.

下面分析上述算法的时间复杂度.

执行步骤 1 的时间复杂度为 $O(|U|)$, 执行步骤 2 的时间复杂度为 $O(|U||C|)$, 执行步骤 4 的时间复杂度为 $O(|U||C|(|C|-1))$, 其余步骤的时间复杂度为常数.

因此, ICBCE 算法的时间复杂度为

$$O(|U|)+O(|U||C|)+O(|U||C|(|C|-1))=O(|U||C|^2).$$

5.2　实例分析

下面我们通过一个关于气象信息的决策表(如表 1 所示) 来分析算法 ICBCE 的有效性. 表 1 中条件属性集 $C=\{a_1, a_2, a_3, a_4\}$ (a_1 表示 outlook, a_2 表示 temperature, a_3 表示 humidity, a_4 表示 windy), 决策属性集 $D=\{d\}$.

经计算可得

$U/C=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7,16\},\{8\},\{9\},\{10\},\{11\},\{12,15\},\{13\},\{14\}\}$,

$U/D=\{\{1,2,6,8,14,16\},\{3,4,5,7,9,10,11,12,13,15\}\}$,

$U/(C-\{a_1\})=\{\{1,3\},\{2\},\{4,8\},\{5,9\},\{6,7,16\},\{10\},\{11,12,15\},\{13\},\{14\}\}$,

$U/(C-\{a_2\})=\{\{1,8\},\{2\},\{3\},\{4\},\{5,10\},\{6\},\{7,12,15,16\},\{9\},\{11\},\{13\},\{14\}\}$,

$U/(C-\{a_3\})=\{\{1\},\{2\},\{3,13\},\{4,10\},\{5\},\{6\},\{7,16\},\{8\},\{9\},\{11\},\{12,15\},\{14\}\}$,

$U/(C-\{a_4\})=\{\{1,2\},\{3\},\{4,14\},\{5,6\},\{7,16\},\{8\},\{9\},\{10\},\{11\},\{12,15\},\{13\}\}$.

从而

$$E_U(D|C)=\frac{2}{256},\ E(D|(C-\{a_1\}))=\frac{8}{256},\ E(D|(C-\{a_2\}))=\frac{6}{256},$$

$$E(D|(C-\{a_3\}))=\frac{2}{256},\ E(D|(C-\{a_4\}))=\frac{6}{256}.$$

因此, $CORE_D^I(C)=\{a_1, a_2, a_4\}$.

表 1　关于气象信息的决策表

U	outlook	temperature	humidity	windy	d
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	Normal	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N
15	Overcast	Mild	Normal	True	P
16	Overcast	Cool	Normal	True	N

表 2 新加入的对象

U	outlook	temperature	humidity	windy	d
17	Sunny	Cool	High	False	N

将新对象 17(如表2 所示) 加入表 1, 并利用算法 ICBCE 求解新的属性核.

由 U/D 得 $Y_p = \{1, 2, 6, 8, 14, 15, 16\} \cup \{17\}$, 则

$$U \cup \{17\}/D = \{\{1, 2, 6, 8, 14, 16, 17\}, \{3, 4, 5, 7, 9, 10, 11, 12, 13, 15\}\}.$$

由 U/C 得 $X_q = \{17\}$, 则

$$U \cup \{17\}/C = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7, 16\}, \{8\}, \{9\}, \{10\}, \{11\}, \{12, 15\}, \{13\}, \{14\}, \{17\}\},$$

$$E_{U\cup\{17\}}(D \mid C) = \frac{1}{(17)^2}\left((16)^2 \times \frac{2}{256} + 2 \times 0\right) = \frac{2}{289}.$$

由 $U/(C - \{a_1\})$ 得 $Z_s = \{17\}$, 则

$$U \cup \{17\}/(C - \{a_1\}) = \{\{1, 3\}, \{2\}, \{4, 8\}, \{5, 9\}, \{6, 7, 16\}, \{10\}, \{11, 12, 15\}, \{13\}, \{14\}, \{17\}\},$$

$$E_{U\cup\{17\}}(D \mid (C - \{a_1\})) = \frac{1}{(17)^2}\left((16)^2 \times \frac{8}{256} + 2 \times 0\right) = \frac{8}{289}.$$

由 $U/(C - \{a_2\})$ 得 $Z_s = \{1, 8\} \cup \{17\}$, 则

$$U \cup \{17\}/(C - \{a_2\}) = \{\{1, 8, 17\}, \{2\}, \{3\}, \{4\}, \{5, 10\}, \{6\}, \{7, 12, 15, 16\}, \{9\}, \{11\}, \{13\}, \{14\}\},$$

$$E_{U\cup\{17\}}(D \mid (C - \{a_2\})) = \frac{1}{(17)^2}\left((16)^2 \times \frac{6}{256} + 2 \times 0\right) = \frac{6}{289}.$$

由 $U/(C - \{a_3\})$ 得 $Z_s = \{9\} \cup \{17\}$, 则

$$U \cup \{17\}/(C - \{a_3\}) = \{\{1\}, \{2\}, \{3, 13\}, \{4, 10\}, \{5\}, \{6\}, \{7, 16\}, \{8\}, \{9, 17\}, \{11\}, \{12, 15\}, \{14\}\},$$

$$E_{U\cup\{17\}}(D \mid (C - \{a_3\})) = \frac{1}{(17)^2}\left((16)^2 \times \frac{2}{256} + 2 \times 1\right) = \frac{4}{289}.$$

由 $U/(C - \{a_4\})$ 得 $Z_s = \{17\}$, 则

$$U \cup \{17\}/(C - \{a_4\}) = \{\{1, 2\}, \{3\}, \{4, 14\}, \{5, 6\}, \{7, 16\}, \{8\}, \{9\}, \{10\}, \{11\}, \{12, 15\}, \{13\}, \{17\}\},$$

$$E_{U\cup\{17\}}(D \mid (C - \{a_4\})) = \frac{1}{(17)^2}\left((16)^2 \times \frac{6}{256} + 2 \times 0\right) = \frac{6}{289}.$$

根据定理 2 可得

$$CORE'_D(C) = \{a_1, a_2, a_3, a_4\}.$$

如果用文献[6]中的算法计算, 可得表 1 加入新对象 17 后的新属性核

$$CORE'_D(C) = \{a_1, a_2, a_3, a_4\}.$$

因此, 对于表 1 本文中的算法与文献[6]中的算法求得的属性核是一致的.

6 结语

通过分析决策表中论域更新时决策属性关于条件属性的条件熵的变化机制, 本文建立了其变化的定量关系, 并提出了一种基于条件熵的增量核求解算法. 算法中计算新的条件熵时对于新对象加入后决策表的不同变化情况都采用了统一的计算公式, 只需找到与新对象属性值相等的条件类和决策类就可以计算出新的条件熵, 进而得到信息观下新决策表的属性核. 该方法可用于在信息观下增量属性核的动态更新, 为海量数据的决策提供了一种有效的处理技术.

参考文献:

[1] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Boston: Kluwer Academic Publishers, 1991.

[2] Liang J Y, Xu Z B. The algorithm on knowledge reduction in incomplete information systems[J]. International Journal of Uncertainty, Fuzziness and System Knowledge Based, 2002, 24(1): 95-103.

[3] 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.
Liang J Y, Qu K S, Xu Z B. Reduction of attribute in information systems[J]. Systems Engineering - Theory & Practice, 2001, 21(12): 76-80.

[4] Skowron A, Rauszer C. The discernibility matrixes and functions in information systems[C]// Slowinski R. Intelligent Decision Support, Handbook of Applications and Advaces of Rough Sets Theory. Dordrecht: Kluwer Academic Publishers, 1992, 331-362.

[5] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligences, 1998, 105(1-2): 77-103.

[6] 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.
Wang G Y. Calculation methods for core attributes of decision table[J]. Chinese Journal of Computers, 2003, 26(5): 611-615.

[7] Wang G Y, Zhao J, An J J, et al. Theoretical study on attribute reduction of rough set theory: Comparison of algebra and information views[C]//Proceedings of the 3rd IEEE International Conference on Cognitive Informatics, Victoria, Canada, 2004, 148-155.

[8] Wang G Y, Yu H, Yang D C. Algebra view and information view of rough sets theory[C]//Proceedings of SPIE, 2001, 4384: 200-207.

[9] 王国胤. 不相容决策信息系统属性核的研究[J]. 上海交通大学学报, 2004, 38(12): 2094-2098.
Wang G Y. Attribute core of inconsistent decision information systems[J]. Journal of Shanghai Jiaotong University, 2004, 38(12): 2094-2098.

[10] Hu X H, Cercone N. Learning in relational databases: A rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338.

[11] 叶东毅, 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.
Ye D Y, Chen Z J. A new discernibility matrix and the computation of a core[J]. Acta Electronica Sinica, 2002, 30(7): 1086-1088.

[12] 杨明, 孙志挥. 改进的差别矩阵及其求核方法[J]. 复旦学报(自然科学版), 2004, 43(5): 865-868.
Yang M, Sun Z H. Improvement of discernibility matrix and the computation of a core[J]. Journal of Fudan University, 2004, 43(5): 865-868.

[13] 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.
Yang M. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix[J]. Chinese Journal of Computers, 2006, 29(3): 407-413.

[14] 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.
Liang J Y, Li D Y. The Uncertainty and Knowledge Acquiring in Imformation Systems[M]. Beijing: Science Press, 2005.

[15] Liang J Y, Chin K S, Dang C Y, et al. A new method for measuring uncertainty and fuzziness in rough set theory[J]. International Journal of General Systems, 2002, 31(4): 331-342.