分布式数据环境中元信息机制在 VPRSM 中的应用①

Applying Meta-Information Mechanism to VPRSM in Distributed Data Environment

张 泳 苏 健 (浙江大学城市学院 网络与计算重点实验室 浙江 杭州 310015)

摘 要: 变精度粗糙集模型(VPRSM)是粗糙集理论的一种扩展。要在分布式数据环境中使用 VPRSM 方法,数据整合是一个代价很高的过程。因此,本文引入了元信息机制,并将元信息整合作为替代数据整合的方法。通过分析与证明,VPRSM 的许多概念可以在元信息机制下等价定义,大量的 VPRSM 的现有方法能够在元信息基础上重新实现,并利用元信息机制降低运算代价。

关键词: 元信息机制 变精度粗糙集模型 数据整合 分布式数据环境

1 引言

粗糙集理论已经成为用于数据挖掘和知识发现的一个重要工具[1-3],可变精度粗糙集模型(VPRSM)是粗糙集理论的一种扩展[4,5]。虽然粗糙集理论被证明是一个很好的工具,但是其中仍有许多问题亟待解决,比如粗糙集方法的有效实现。许多现有的粗糙集方法都假设数据是处于集中状况下的,但在信息化时代,数据通常分布在不同的位置,为了使用这些方法,就必须付出巨大的代价去进行数据整合。

本文首先介绍分布式数据环境中的元信息机制及元信息整合方法,然后探讨元信息机制与 VPRSM 的关系,并在元信息机制的基础上,对许多 VPRSM 的概念进行等价定义。这些等价定义表明,许多现存 VPRSM 方法可在元信息基础上高效实现。

2 信息系统与元信息

粗糙集理论将所处理的数据表示为信息系统。信息系统可表示为 $S=<U, A>$,其中 $U\neq\varnothing$ 是一个对象空间,$A=C\cup D$ 是一个属性集,C 为条件属性集,D 为决策属性集。在分布式数据环境中,信息系统被扩展为分布式信息系统,它由一组两两不相交的子系统组成。令对象空间 $U=U_1\cup U_2\cup\cdots\cup U_m$,其中如果 $i\neq j$,则 $U_i\cap U_j=\Phi$,那么 $S_i=<U_i,A>$ 就表示第 i 个子系统[6,7]。

根据属性集 $B\subseteq A$,可对对象空间 U 进行划分,记为 $B^*=U/B$。$d=(x,|[x]_B|)$ 称为类描述符,用于描述一个等价类 $[x]_B\in B^*$,其中 x 是特征元素,$|[x]_B|$ 是类的基数。如果 $B\subseteq C$,则 $(x,|[x]_B|)$ 被称为条件类描述符,同样的,如果 $B\subseteq D$,$(x,|[x]_B|)$ 被称为决策类描述符。

如果 $I=<M_{CD},F_C,F_D>$,表示信息系统 $S=<U,A>$ 的元信息,那么其中:

(1) $F_C=\{(x_1,|[x_1]_C|),(x_2,|[x_2]_C|),\cdots,(x_k,|[x_k]_C|)\}$,$k=|C^*|$

(2) $F_D=\{(y_1,|[y_1]_D|),(y_2,|[y_2]_D|),\cdots,(y_l,|[y_l]_D|)\}$,$l=|D^*|$

(3) $M_{CD}=(m_{ij})$ 是一个类矩阵,$m_{ij}=|[x_i]C\cap[y_j]D|$

任何一个类描述符都与三个参数有关:class(d)、ce(d) 和 card(d),这些参数分别表示等价类、特征元素及类基数。同样的,m(p,q) 表示与 $p\in F_C$、$q\in F_D$ 有关的类矩阵中的一项。

元信息机制通过类描述符代表一个等价类。由于类描述符仅仅需要等价类的基数以及一个特征元素,其存储空间开销大大降低。因此,元信息使用紧凑方式描述一个信息系统,与原始底层数据相比,其存储代价非常低。

① 基金项目:国家自然科学基金资助(60473097)

3 元信息整合

元信息整合的主要目的是将分布式信息系统的各子系统的元信息整合为信息系统的元信息。元信息整合处理过程不对底层基础数据(即对象空间)直接操作,而是在元信息层面上进行集成操作,因此元信息整合的代价将远小于底层数据的整合代价。元信息整合的任务包括:(1)类描述符的整合;(2)类矩阵的整合。令 $I = <M_{CD}, F_C, F_D>$ 为信息系统 $S = <U, A>$ 的元信息,$I_i = <M^i_{CD}, F^i_C, F^i_D>$ 是第 i 个子系统 S_i 的元信息。元信息整合就是从 $I_1, \ldots, I_i, \ldots, I_n$ 生成 I。

3.1 类描述集整合

类描述集整合包括各个信息子系统元信息中条件类描述集整合以及决策类描述集整合。首先考虑条件类描述集的整合方法,即如何从 $F^1_C, F^2_C, \ldots, F^i_C, \ldots, F^n_C$ 中生成 F_C。令 $Q(C) = \cup \{F^i_C : 1 \le i \le n\}$,即 Q(C) 是所有子系统元信息中条件类描述符集合的并集。对于任意两个条件类描述符 $p, q \in Q(C)$ 根据等价关系 C,如果 p 与 q 的特征元素 ce(p)、ce(q) 是不可辨别的,那么可认为 p 与 q 也是不可辨别的,记为 pCq。令 $Z(p,C) = \{q \in Q(C) : pCq\}$,并记 Q(C)/C 为 Q(C) 根据条件属性集 C 的一个划分。

命题 1:令 $X, Y \in Q(C)/C$,则

(1) $X \ne Y \Rightarrow \forall p \in X \ \forall q \in Y \ [\neg (pCq)]$

(2) $\exists p \in X \ \exists q \in Y \ [pCq] \Rightarrow X = Y$

证略。

命题 2:令 $Z(p,C) = \{q \in Q(C) : pCq\}$,则

(1) $\forall p \in Q(C) \ \exists T \in U/C \ [ce(p) \in T \wedge |T| = \Sigma \{card(q) : q \in Z(p,C)\}]$

(2) $\forall T \in U/C \ \exists p \in Q(C) \ [ce(p) \in T \wedge |T| = \Sigma \{card(q) : q \in Z(p,C)\}]$

证略。

命题 1 说明任何一对不同块的类描述符是可辨别的,而相同块的类描述符是不可辨别的。

命题 2 中的式(1)说明了一种从 Z(p,C) 产生 F_C 的一个条件类描述符的方法,即 $(ce(p), \Sigma \{card(q) : q \in Z(p,C)\})$ 是 U/C 的条件类的一个有效描述符。式(2)证实了 F_C 的所有条件类描述符都可以从 Q(C) 计算得到。因此,我们可以通过以下步骤获得 F_C:(1)获得 Q(C)/C 的划分;(2)从 Q(C)/C 的一个块中计算 F_C 的一个类描述符;(3)重复步骤(2)直至所有的块被操作。

各个信息子系统元信息中决策类描述集的整合方法与上述方法思路一致。令 $Q(D) = \cup \{F^i_D : 1 \le i \le n\}$,可通过类似方法从 Q(D)/D 计算获得 F_D。

3.2 类矩阵整合

前文讨论了信息子系统元信息中条件类描述集以及决策类描述集的整合方法。下面讨论类矩阵的整合方法。前文已述,可从 Q(C)/C 的一个块 $P(d^C)$ 获得一个条件类描述符 $d^C \in F_C$;同样,也可从 Q(D)/D 的一个块 $P(d^D)$ 获得一个决策类描述符 $d^D \in F_D$。对 $p \in P(d^C)$,令 sys(p) 表示 p 所来自的子系统;同样,对 $q \in P(d^D)$,令 sys(q) 表示 q 所来自的子系统。

命题 3:令 $P(d^C) \in Q(C)/C$, $P(d^D) \in Q(D)/D$,其中 $d^C \in F_C, d^D \in F_D$。$m(d^C, d^D)$ 表示类矩阵 M_{CD} 的项,则 $m(d^C, d^D) = \Sigma \{m^k(p,q) : p \in P(d^C) \wedge q \in P(d^D) \wedge sys(p) = sys(q) = k\}$,其中 $m^k(p,q)$ 表示信息系统 S 第 k 个子系统的类矩阵第 p, q 项。

证明:令 $P(d^C) = \{p_1, p_2, \cdots, p_t\}$,其中 $t = |P(d^C)|$,那么 $class(d^C) = class(p_1) \cup class(p_2) \cup \cdots \cup class(p_t)$。

同样,令 $P(d^D) = \{q_1, q_2, \cdots, q_g\}$,其中 $g = |P(d^D)|$,那么,$class(d^D) = class(q_1) \cup class(q_2) \cup \cdots \cup class(q_g)$。根据类矩阵定义,$m(d^C, d^D) = |class(d^C) \cap class(d^D)| = \Sigma \{|class(p_i) \cap class(q_j)| : 1 \le i \le t, 1 \le j \le g\}$。

如果 p_i 与 q_j 来自同一子系统,即 $sys(p_i) = sys(q_j) = k$,那么 $|class(p_i) \cap class(q_j)| = m^k(p_i, q_j)$。

如果 p_i 与 q_j 来自不同子系统,即 $sys(p_i) \ne sys(q_j)$,那么 $|class(p_i) \cap class(q_j)| = 0$。

因此,$m(d^C, d^D) = \Sigma \{m^k(p,q) : p \in P(d^C) \wedge q \in P(d^D) \wedge sys(p) = sys(q) = k\}$。

命题 3 显示了整合类矩阵的一种方法,可以通过以下步骤获得 M_{CD}:(1)为所有的 $d^C \in F_C$ 获得 $P(d^C)$;(2)为所有的 $d^C \in F_D$ 获得 $P(d^D)$;(3)获得一个类矩阵项 $m(d^C, d^D)$;(4)重复步骤(3)直至所有的 d^C 和 d^D 被操作。

4　元信息基础上的 VPRSM 概念

4.1 决策类的 β 近似

在 VPRSM 中,集合包含关系被扩展至近似包含关系,以允许一定程度的错误分类。令 $X\subseteq U, Y\subseteq U$,其中 $X\neq\Phi, Y\neq\Phi, \beta(0<\beta<0.5)$ 是一个预定义的误差因子,则近似包含关系可定义为:

$X\subseteq_\beta Y$ 当且仅当 $c(X,Y)\leq\beta$

其中 $c(X,Y)$ 是错误分类的度量,定义为:

$$c(X,Y)=\begin{cases}1-|X\cap Y|/|X| & \text{if } |X|>0\\ 0 & \text{if } |X|=0\end{cases}$$

VPRSM 将集合近似概念推广为 β 近似。令 $R\subseteq A$, $R^*=U/R$,则 $X\subseteq U$ 的 β 下近似可被定义为:

$$\underline{R}_\beta X=\cup\{E\in R^*: c(E,X)\leq\beta\}$$

$X\subseteq U$ 的 β 上近似可被定义为:

$$\overline{R}_\beta X=\cup\{E\in R^*: c(E,X)<1-\beta\}$$

$X\subseteq U$ 的 β 边界可被定义为:

$$BNR_\beta X=\cup\{E\in R^*:\beta<c(E,X)<1-\beta\}$$

元信息机制主要考虑的是决策类描述符的近似。对于类描述 $d\in F_C$ 或 F_D,令 $class(d)$ 表示 d 所代表的等价类。令 $m(d^C,d^D)=|class(d^C)\cap class(d^D)|$ 表示类矩阵 M_{CD} 中由 d^C, d^D 决定的项。对于条件属性子集 $R\subseteq C$,C^* 的几个条件类可以被合并成一个 R^* 的新类。令 $R(d)=\{q\in F_C: q\ R\ d\}$,其中 $d\in F_C$。显然,$\cup\{class(q): q\in R(d)\}$ 就是 R^* 中的一个类。

命题 4 如果 $Y=class(d^D)$,其中 $d^D\in F_D$,且 $X=\cup\{class(q): q\in R(d^C)\}$,其中 $d^C\in F_C$,那么

$X\subseteq_\beta Y\Leftrightarrow t(d^C,d^D,R)\leq\beta$,其中 $t(d^C,d^D,R)=1-\Sigma\{m(p,d^D): p\in R(d^C)\}/\Sigma\{card(p): p\in R(d^C)\}$。

证明:令 $R(d^C)=\{q_1,q_2,\cdots,q_t\}$,其中 $t=|R(d^C)|$. 令 $X=class(q_1)\cup class(q_2)\cdots\cup class(q_t)$,那么 $|X\cap Y|=|(class(q_1)\cup class(q_2)\cdots\cup class(q_t))\cap class(d^D)|=\Sigma\{|class(q_i)\cap class(d^D)|: 1\leq i\leq t\}$. 由于 $|class(q_i)\cap class(d^D)|=m(q_i,d^D)$,$|X\cap Y|=\Sigma\{m(p,d^D): p\in R(d^C)\}$,显然,$|X|=\Sigma\{|class(q_i)|: 1\leq i\leq t\}$,所以 $|X|=\Sigma\{card(p): p\in R(d^C)\}$,因此,$X\subseteq_\beta Y\Leftrightarrow t(d^C,d^D,R)\leq\beta$。

命题 4 说明,能够直接通过元信息的类矩阵和类描述符验证一个条件类是否 β 包含于一个决策类。令 $R\subseteq C$ 是一个条件属性集,则 $d^D\in F_D$ 所描述的决策类的 β 下近似可定义为:

$$\underline{R}_\beta d^D=\cup\{R(d^C): d^C\in F_C\wedge t(d^C,d^D,R)\leq\beta\}.$$

β 上近似可定义为:

$$\overline{R}_\beta d^D=\cup\{R(d^C): d^C\in F_C\wedge t(d^C,d^D,R)<1-\beta\}.$$

β 边界可定义为:

$$BNR_\beta d^D=\cup\{R(d^C): d^C\in F_C\wedge\beta<t(d^C,d^D,R)<1-\beta\}.$$

根据 VPRSM,决策类 Y 的 β 近似度被定义为:

$$\alpha(R,\beta,Y)=card(\underline{R}_\beta Y)/card(\overline{R}_\beta Y)$$

在元信息机制中,$d^D\in F_D$ 所描述的决策类的 β 近似度可定义为:

$$\alpha(R,\beta,d^D)=\Sigma\{card(p): p\in\underline{R}_\beta d^D\}/\Sigma\{card(q): q\in\overline{R}_\beta d^D\}$$

因此,决策类的 β 精确近似也能够通过元信息直接计算得到。

4.2 β 依赖与 β 约简

在 VPRSM 中,划分 Q^* 的 β 正区域 $POS(P,Q,\beta)$ 被定义为:

$POS(P,Q,\beta)=\cup\{\underline{B}_\beta(Y): Y\in Q^*\}$,其中 $P\subseteq A$, $Q\subseteq A$, $\underline{B}_\beta(Y)=\{X: X\in B^*\wedge X\subseteq_\beta Y\}$。

划分 Q^* 的 β 正区域包括允许 β 误差率前提下的所有可被正确分类的对象。

属性集 Q 对属性集 P 的 β 依赖度 $\gamma(P,Q,\beta)$ 定义为:$\gamma(P,Q,\beta)=|POS(P,Q,\beta)|/|U|$。元信息机制重点考虑决策属性集 D 对条件属性集 $B\subseteq C$ 的 β 依赖 $\gamma(B,D,\beta)$,显然 $\gamma(B,D,\beta)=|POS(B,D,\beta)|/|U|$。

由于条件属性集 C 中存在某些属性,这些属性对于决策是冗余的。在允许一定程度误差 β 前提下,消除这些冗余属性所得到的条件属性集合,被称为条件属性集的 β 约简。显然,β 约简是条件属性集的一个子集。令 $RED(C,D,\beta)$ 表示 β 约简,需要同时满足下面两个条件:

(1) $\gamma(RED(C,D,\beta),D,\beta)=\gamma(C,D,\beta)$

(2) $\forall a\in RED(C,D,\beta): \gamma(RED(C,D,\beta)-\{a\},D,\beta)\neq\gamma(C,D,\beta)$

β 约简保留了 D 对 C 的 β 依赖度，并且，β 约简也保留了整个条件属性集 C 的分类能力。

命题 5 令 $B \subseteq C$，那么

(1) $POS(B,D,\beta) \subseteq POS(C,D,\beta)$

(2) $\gamma(B,D,\beta) = \gamma(C,D,\beta) \Leftrightarrow |POS(B,D,\beta)| = |POS(C,D,\beta))|$

证明：(1) 根据 β 正域定义，显然有 $POS(B,D,\beta) \subseteq POS(C,D,\beta)$。

(2) 根据 β 依赖定义，有 $\gamma(B,D,\beta) = \gamma(C,D,\beta) \Rightarrow |POS(B,D,\beta)| = |POS(C,D,\beta))|$。令 $X \in C^*$，那么 $X \subseteq POS(B,D,\beta) \Rightarrow X \subseteq POS(C,D,\beta))$，所以 $POS_B(D) \subseteq POS_C(D)$，因此 $|POS_B(D)| = |POS_C(D)| \Rightarrow POS_B(D) = POS_C(D)$。

命题 6 令 $B \subseteq C$，则 $|POS(B,D,\beta)| = \Sigma \{card(q) : q \in Z\}$，其中 $Z = \cup \{\underline{R}_\beta d^D : d^D \in F_D\}$

证明 根据 β 正域定义，$POS(B,D,\beta) = \cup \{\underline{B}_\beta(Y) : Y \in D^*\}$，所以 $|POS(B,D,\beta)| = \Sigma \{|\underline{B}_\beta(Y)| : Y \in D^*\}$。由于每个 $Y \in D^*$ 对应某一类描述符 $d^D \in F_D$，所以 $\underline{B}_\beta(Y)$ 对应 $\underline{R}_\beta d^D$，其中 $\underline{R}_\beta d^D = \cup \{R(d^C) : d^C \in F_C \wedge t(d^C,d^D,R) \leq \beta\}$，因而 $|\overline{B}_\beta(Y)| = \Sigma \{card(q) : q \in R(d^C)\}$。因此，$|POS(B,D,\beta)| = \Sigma \{|\underline{B}_\beta(Y)| : Y \in D^* = \Sigma \{card(q) : q \in Z\}$，其中 $Z = \cup \{\underline{R}_\beta d^D : d^D \in F_D\}$。

从命题 6 可知，划分 D^* 的 β 正区域基数能够通过元信息获得。按照 β 约简和 β 依赖的定义，β 正区域基数计算是进行 β 约简的重要步骤，因此，我们可以通过元信息获得 β 约简，并不需要依赖整个对象空间 U。由于 β 正区域基数计算的运算代价远小于 β 正区域的计算，基于元信息的 β 约简方法的开销将大大降低。

5 结论

本文讨论了在分布式数据环境中元信息机制中元信息整合方法，以及在 VPRSM 中应用的关键问题。在元信息机制中，由于元信息只处理类描述符和类矩阵，存储和计算开销都明显降低，通过元信息整合取代底层基础数据整合，将大大减低 VPRSM 方法的运算代价。另外，分析表明，在元信息机制的基础上，一些 VPRSM 的重要概念可以被等价定义。因此，现存的许多 VPRSM 方法将可以在元信息基础上重新实现，并利用元信息机制降低运算代价。

参考文献

1 Z. Pawlak. Rough set: Theoretical Aspects of Reasoning About Data. Kluwer Academic. Dordrecht, The Netherlands, 1991.

2 Z. Pawlak, A. Skowron. Rough sets: Some extensions. Information Sciences, 2007, 177(1): 28 – 40.

3 J. Komorowski, Z. Pawlak, L. Polkowski, et al. Rough sets: A Tutorial. In: Pal S. K., A. Skowron, eds., Rough fuzzy hybridization. A new trend in decision – making. Springer, 1999. 3 – 98.

4 W. Ziarko. variable precision rough set model. Journal of Computer and System Sciences, 1993. 39 – 59.

5 张贤勇，莫智文. 变精度粗糙集. 模式识别与人工智能，2004，(6).

6 J. Su, J. Gao, Methods for meta – information maintenance in rough set theory. Proceedings of 2003 International Conference on Machine Learning and Cybernetics, November 2 – 5 2003, Xi'an, China, pp. 1527 – 1532.

7 苏健，高济. 基于元信息的粗糙集规则并行挖掘方法. 计算机科学，2003，30(3).