

基于不完备区间值信息系统的决策粗糙集^{*}

张鑫, 李续武, 路艳丽, 陈玉金

(空军工程大学 防空反导学院, 西安 710043)

摘要: 在不完备区间值信息系统中, 提出一种基于极大相容类的决策粗糙集模型。针对不完备区间值信息系统中属性相似度存在的缺陷, 对属性相似度进行改进。在不完备区间值信息系统中, 由于容差关系下建立粗糙集模型存在冗余度高、分类精度低的问题, 采用极大相容类代替等价类, 结合贝叶斯最小风险决策原则, 建立决策粗糙集模型。经证明, 基于极大相容类建立粗糙集模型可有效提高分类精度。最后, 基于正域分布不变的原则提出基于区分矩阵的属性约简算法并将该算法应用于实例。

关键词: 不完备区间值信息系统; 属性相似度; 决策粗糙集; 区分矩阵

中图分类号: TP18

文献标志码: A

文章编号: 1001-3695(2017)01-0110-04

doi:10.3969/j.issn.1001-3695.2017.01.023

Decision-theoretic rough set based on incomplete interval-valued information system

Zhang Xin, Li Xuwu, Lu Yanli, Chen Yujin

(Air & Missile Defence College, Air Force Engineering University, Xi'an 710043, China)

Abstract: In incomplete interval-valued information system, this paper proposed the decision-theoretic rough set based on maximal consistent class. Considering the insufficient about attribute similarity in incomplete interval-valued information system, it provided the improved attribute similarity. Then, in order to solve the model's high redundancy and low classification accuracy in the information system, this paper replaced equivalence class with the maximal consistent class and set up the decision-theoretic rough set model combined with Bayesian smallest risk theory. And this paper proved that set up the model based on maximal consistent class can improve the classification accuracy. Finally, this paper proposed the attribute reduction algorithm based on indiscernibility matrix and remains distribution of positive region and applied it to a case.

Key words: incomplete interval-valued information system; attribute similarity; decision-theoretic rough set; indiscernibility matrix

0 引言

现实生活中信息系统通常是复杂多样的, 很多测量值都表示为在一定范围内的连续区间值; 而且由于数据测量的误差、对数据的理解或获取的限制等众多原因, 使得所面临的信息系统往往是不完备的。经典粗糙集理论^[1]是一种处理不确定性、模糊性问题的数学工具, 是经典集合理论的推广。但经典粗糙集理论主要对分类数据或离散值进行操作, 因此基于不完备信息系统的粗糙集理论得到了广泛的研究^[2-4]。其中, 以不完备区间值信息系统 IIS(incomplete interval-valued information system)为研究对象, 文献[5]结合集对分析中联系度的概念提出相似联系度容差关系; 文献[6]提出了基于广义优势关系的粗糙集模型, 并对分类质量的定义进行了改进; 文献[7]对容差关系进行拓展, 提出基于相似度和相似率的双精度容差关系。

Yao 在长期研究概率粗糙集过程中, 考虑到概率粗糙集模型利用两个参数将论域分为三个区域, 即正域、负域和边界域, 提出符合人类实际认知能力的三支决策模式, 并结合贝叶斯最小风险决策原则给出了概率粗糙集中两个参数在现实中的语义解释, 从而建立决策粗糙集模型^[8,9]。决策粗糙集理论的提

出为粗糙集理论在实际问题中的应用建立起一座桥梁, 基于该理论的研究已逐渐成为热点。Liu 等人讨论了基于直觉模糊数、三角模糊数的决策粗糙集模型^[10], Yao 等人对决策粗糙集的属性约简问题进行了研究^[11], Jia 等人基于风险损失最小的原则对决策粗糙集的属性约简进行了研究^[12]。

在 IIS 中通常采用属性相似度来衡量两个对象关于某个属性的相似程度, 文献[5]中的属性相似度定义: 若两个对象中关于某属性的属性值中存在缺失值, 则两个对象关于该属性的相似度为 1, 显然这样的定义存在片面性。因此本文结合属性的值域对属性相似度的定义进行改进并进一步提出基于改进属性相似度的容差关系。由于基于容差关系进行分类存在冗余度高、分类粒度粗等问题, 本文基于极大相容类建立决策粗糙集模型。文章最后给出基于区分矩阵的属性约简算法, 并将该算法应用于实例。

1 决策粗糙集

称 $IS = (U, AT, V, f)$ 为信息系统, 其中 U 是非空有限的对象集合; AT 是非空有限的属性集合; f 为信息函数, $\forall x \in U$, $f(x, a)$ 表示对象 x 在属性 a 上的取值; V 为全体属性的值域。

收稿日期: 2015-12-09; 修回日期: 2016-01-20 基金项目: 国家自然科学基金资助课题(61272011)

作者简介: 张鑫(1992-), 男(蒙古族), 内蒙古巴彦淖尔人, 硕士, 主要研究方向为粗糙集与智能信息处理(Z_x92kgd@163.com); 李续武(1966-), 男, 河南濮阳人, 教授, 硕导, 主要研究方向为粗糙集与智能信息处理等; 路艳丽(1980-), 女, 陕西大荔人, 博士, 主要研究方向为智能信息处理等; 陈玉金(1992-), 男, 辽宁兴城人, 硕士, 主要研究方向为粗糙集与智能信息处理。

定义 1^[1] 给定信息系统 IS , 其中属性子集 $A \subseteq AT$, 可构成不可分辨关系 $IND(A)$:

$$IND(A) = \{(x, y) \in U^2 : f(x, a) = f(y, a), \forall a \in A\} \quad (1)$$

$IND(A)$ 是自反、对称、传递的, 因此 $IND(A)$ 是一个等价关系。

记 $[x]_A = \{y \in U : (x, y) \in IND(A)\}$ 是对象 x 在等价关系 A 下的等价类, 则决策粗糙集近似算子可定义如下。

定义 2^[8] 给定信息系统 IS , 其中属性子集 $A \subseteq AT$, 则对于 $X \subseteq U$ 的决策粗糙集下近似与上近似分别为

$$\begin{aligned} \underline{\text{apr}}_{(\alpha, \beta)}(X) &= \{x \in U | P(X | [x]_A) \geq \alpha\} \\ \overline{\text{apr}}_{(\alpha, \beta)}(X) &= \{x \in U | P(X | [x]_A) > \beta\} \end{aligned} \quad (2)$$

其中, $P(X | [x]_A) = \frac{|[x]_A \cap X|}{|[x]_A|}$ 为条件概率。

设行为集为 $\{a_p, a_b, a_n\}$, 分别表示接受某事物采取的决策行为、待定的决策行为、拒绝某事物采取的决策行为, 即分别确定 $x \in \text{pos}(X)$ 、 $x \in \text{bn}(X)$ 、 $x \in \text{neg}(X)$; 设 λ_{pp} 、 λ_{bp} 、 λ_{np} 分别为 $x \in X$ 时采取行为 a_p 、 a_b 和 a_n 所产生的风险损失, 相应地, λ_{pn} 、 λ_{bn} 、 λ_{nn} 分别为 $x \in \neg X$ 时采取行为 a_p 、 a_b 和 a_n 所产生的风险损失。则根据决策粗糙集理论, 式(2)中 α 、 β 由风险损失分别表示为

$$\alpha = \frac{\lambda_{pn} - \lambda_{bn}}{\lambda_{bp} - \lambda_{pp} + \lambda_{pn} - \lambda_{bn}}, \beta = \frac{\lambda_{bn} - \lambda_{nn}}{\lambda_{np} - \lambda_{bp} + \lambda_{bn} - \lambda_{nn}} \quad (3)$$

显然, 经典决策粗糙集是建立在等价关系的基础上的, 因此它只能处理具有离散属性值的完备信息系统。然而现实环境中测得的数据往往是在一定范围内变化的连续值, 如时间、功率、温度等, 这些值的范围常常采用区间数据进行描述, 因此研究基于不完备区间值信息系统的决策粗糙集理论具有必要性。

2 基于 IIS 的决策粗糙集

称 $K = (U, A, V, F)$ 是一个信息系统, 其中 $U = \{x_1, x_2, \dots, x_n\}$ 为有限对象集, 称为论域; $A = \{a_1, a_2, \dots, a_m\}$ 为有限属性集; F 为 U 与 A 之间的关系集, 即 $F = \{f_k : U \rightarrow V_k (k \leq m)\}$, 其中 V_k 为 a_k 的值域。若 $f(x_i, a_k) = [l_i^k, r_i^k]$, $l_i^k \leq r_i^k$, 则称 K 为区间值信息系统。若存在 $x_i \in U, a_k \in A, f(x_i, a_k) = *$ (“*”表示缺失值), 则表明对象 x_i 在属性 a_k 上的值未知, 此时称 K 是不完备区间值信息系统 IIS。

显然, 将经典决策粗糙集理论直接拓展到 IIS 较为困难, 因此采用 IIS 中基于属性相似度定义的容差关系代替经典粗糙集中的等价关系, 在 IIS 下对决策粗糙集理论进行拓展。

2.1 基于 IIS 的属性相似度

定义 3^[5] 设 $K = (U, A, V, F)$ 是不完备区间值信息系统, $a_k \in A, x_i, x_j$ 为 U 中的任意两个对象, 当这两个对象在 a_k 下的属性值均不为缺失值时, 即 $f(x_i, a_k) = [l_i^k, r_i^k], f(x_j, a_k) = [l_j^k, r_j^k]$, 称

$$S_{ij}^k = \frac{|[l_i^k, r_i^k] \cap [l_j^k, r_j^k]|}{|[l_i^k, r_i^k] \cup [l_j^k, r_j^k]|} \quad (4)$$

为对象 x_i 和 x_j 关于属性 a_k 的相似度。其中 $|\cdot|$ 在点集中表示集合所含元素个数, 这里表示闭区间的长度, 并规定空集与单点集的区间长度是零。当这两个属性值均为单值, 若相等则相似度为 1, 不相等则相似度为 0。若这两个属性值至少有一个为缺失值时, 规定它们之间的相似度为 1。

由定义 3 可知, 当两个对象中关于属性 a_k 的属性值中存

在缺失值时, 则两个对象关于属性 a_k 的相似度为 1。由式(4)可知, 两个对象的属性值均为区间值时, 关于属性 a_k 的相似度只有在两个对象的属性值为同一区间时为 1。由于笔者认为信息表中属性 a_k 的未知属性值为“遗漏”型, 所以未知的属性值可能取该属性值域范围内任意值, 即 $f(x_i, a_k) = * = [\min V_k, \max V_k]$ 。定义 3 中两个对象关于属性 a_k 的相似度显然具有一定的局限性, 于是对相似度定义进行改进如下。

定义 4 设 $K = (U, A, V, F)$ 是不完备区间值信息系统, $a_k \in A, x_i, x_j$ 为 U 中的任意两个对象, 规定空集与单点集的区间长是零, 当两个对象关于属性 a_k 的属性值均不为单值时, 对象 x_i 和 x_j 关于属性 a_k 的相似度定义为

$$s_{ij}^k = \begin{cases} \frac{|[l_i^k, r_i^k] \cap [l_j^k, r_j^k]|}{|[l_i^k, r_i^k] \cup [l_j^k, r_j^k]|}, & (f(x_i, a_k) = [l_i^k, r_i^k]) \wedge (f(x_j, a_k) = [l_j^k, r_j^k]) \\ \frac{|[l_i^k, r_i^k]|}{|[\min V_k, \max V_k]|}, & (f(x_i, a_k) = [l_i^k, r_i^k]) \wedge (f(x_j, a_k) = *) \\ \frac{|[l_j^k, r_j^k]|}{|[\min V_k, \max V_k]|}, & (f(x_i, a_k) = *) \wedge (f(x_j, a_k) = [l_j^k, r_j^k]) \\ 1, & (f(x_i, a_k) = *) \wedge (f(x_j, a_k) = *) \end{cases} \quad (5)$$

当这两个属性值均为单值, 若相等则相似度为 1, 不相等则相似度为 0。

2.2 基于 IIS 的极大相容类

根据定义 4 可以给出 IIS 下容差关系及容差类定义如下。

定义 5 给定不完备区间值信息系统 $K = (U, A, V, F)$, 对 $a_k \in A, \omega$ 是相似度阈值, 在不完备区间值信息系统上定义如下容差关系:

$$\Gamma_A^\omega = \{(x_i, x_j) \in U \times U | s_{ij}^k \geq \omega, \forall a_k \in A\} \quad (6)$$

显然 Γ_A^ω 满足自反性、对称性, 但不满足传递性。

记 $\Gamma_A^\omega(x_i) = \{x_j \in U | (x_i, x_j) \in \Gamma_A^\omega\}$ 为容差类, 下面给出容差关系下集合 X 的下近似和上近似的定义。

定义 6 给定不完备区间值信息系统 $K = (U, A, V, F)$, 对 $a_k \in A, \omega$ 是相似度阈值, $X \subseteq U$, 则 X 的下近似与上近似分别为

$$\begin{aligned} \underline{\Gamma}_A^\omega(X) &= \{x_i \in U | \Gamma_A^\omega(x_i) \subseteq X\} \\ \overline{\Gamma}_A^\omega(X) &= \{x_i \in U | \Gamma_A^\omega(x_i) \cap X \neq \emptyset\} \end{aligned} \quad (7)$$

不完备区间值信息系统 $K = (U, A, V, F)$ 上的容差关系 Γ_A^ω 是自反的, 因此 $\{\Gamma_A^\omega(x_i) : x_i \in U\}$ 构成论域上的一个覆盖。基于容差关系进行分类存在如下问题: 分类粒度粗、数目多、冗余度高; 同一类中的任意两个对象不一定具有相同的属性特性, 误分率高。为了解决以上问题, 提高分类精度, 引入极大相容类的概念。

定义 7 给定不完备区间值信息系统 $K = (U, A, V, F)$, ω 是相似度阈值。若 $M \subseteq U$ 对任意的 $x_i, x_j \in M$ 均满足 $(x_i, x_j) \in \Gamma_A^\omega$, 则称 M 是 K 中关于属性集 A 的相容类。若对任意的 $x_m \in U - M_A^\omega(x_i)$ 必存在 $x_j \in M_A^\omega(x_i)$ 使得 $(x_j, x_m) \notin \Gamma_A^\omega$ 。此时, 称 $M_A^\omega(x_i)$ 为 K 中关于属性集 A 的极大相容类。

定理 1 设 $M_A^\omega(x_i)$ 为区间信息系统 $K = (U, A, V, F)$ 中关于属性集 A 的所有包含对象 x_i 的极大相容类的集合, 则对任意的 $x_i \in U, \forall B \subseteq A$, 有 $\Gamma_B^\omega(x_i) = \cup \{M_B^\omega(x_i) \in m_B^\omega(x_i)\}$ 。

证明: 由定义 7 以及 $\Gamma_B^\omega(x_i) = \{x_j \in U | (x_i, x_j) \in \Gamma_B^\omega\}$ 可知, $\forall x_j \in \Gamma_B^\omega(x_i)$ 存在 $x_j \in \cup \{M_B^\omega(x_i) \in m_B^\omega(x_i)\}$, 即 $\Gamma_B^\omega(x_i) \subseteq \cup \{M_B^\omega(x_i) \in m_B^\omega(x_i)\}$; 又 $\forall x_j \in M_B^\omega(x_i), M_B^\omega(x_i) \in m_B^\omega(x_i)$, 即

x_i, x_j 满足容差关系 Γ_B^ω , 因此 $x_j \in \Gamma_B^\omega(x_i)$, 进而 $\cup \{M_B^\omega(x_i) \in m_B^\omega(x_i)\} \subseteq \Gamma_B^\omega(x_i)$ 。综上所述, $\Gamma_B^\omega(x_i) = \cup \{M_B^\omega(x_i) \in m_B^\omega(x_i)\}$ 。由定义 7 可知, 极大相容类是论域中对象两两相容的最大集合, 任意相容类一定可以拓展为极大相容类; 极大相容类是论域中满足传递关系的最大子集; 极大相容类构成了论域上的一个覆盖。下面给出采用极大相容类 $M_A^\omega(x_i)$ 定义的集合 X 的下近似和上近似。

定义 8 给定不完备区间值信息系统 $K = (U, A, V, F)$, 对 $a_k \in A, \omega$ 是相似度阈值, $X \subseteq U$, 则 X 的下近似与上近似分别为

$$\begin{aligned} \underline{\text{APR}}_A^\omega(X) &= \{x_i \in U \mid M_A^\omega(x_i) \subseteq X\} \\ \overline{\text{APR}}_A^\omega(X) &= \{x_i \in U \mid M_A^\omega(x_i) \cap X \neq \emptyset\} \end{aligned} \quad (8)$$

根据经典粗糙集理论中粗糙集近似精度定义, 粗糙集近似精度为粗糙集下近似的基数与粗糙集上近似的基数之比。

定理 2 给定不完备区间值信息系统 $K = (U, A, V, F)$, ω 是相似度阈值, $\forall B \subseteq A, X \subseteq U$, 则:

$$\frac{|\underline{\text{APR}}_B^\omega(X)|}{|\overline{\text{APR}}_B^\omega(X)|} \geq \frac{|\underline{\Gamma}_B^\omega(X)|}{|\overline{\Gamma}_B^\omega(X)|} \quad (9)$$

证明: 设 $M^\omega(B)$ 为信息系统上属性集 B 所决定的所有极大相容类的集合。由于 $\underline{\Gamma}_B^\omega(X) = \{x_i \in U \mid \Gamma_B^\omega(x_i) \cap X \neq \emptyset\} = \cup \{\Gamma_B^\omega(x_i) \mid x_i \in X\}$, 由定理 1 可知 $\overline{\Gamma}_B^\omega(X) = \cup \{M_B^\omega(x_i) \in m_B^\omega(x_i) \mid x_i \in X\} = \cup \{M_B^\omega(x_i) \in M^\omega(B) \mid M_B^\omega(x_i) \cap X \neq \emptyset\} = \underline{\text{APR}}_B^\omega(X)$; $\forall x_i \in \overline{\Gamma}_B^\omega(X)$, 有 $\Gamma_B^\omega(x_i) \subseteq X$, 由于 $\Gamma_B^\omega(x_i) = \cup \{M_B^\omega(x_i) \in m_B^\omega(x_i)\}$, 故 $M_B^\omega(x_i) \subseteq X$, 从而 $x_i \in \underline{\text{APR}}_B^\omega(X)$, 则 $\overline{\Gamma}_B^\omega(X) \subseteq \underline{\text{APR}}_B^\omega(X)$ 。综上所述, $\frac{|\underline{\text{APR}}_B^\omega(X)|}{|\overline{\text{APR}}_B^\omega(X)|} \geq \frac{|\underline{\Gamma}_B^\omega(X)|}{|\overline{\Gamma}_B^\omega(X)|}$ 。

由定理 2 可以看出, 采用极大相容类对粗糙集下近似与上近似定义较采用改进相似度容差类定义粗糙集下近似与上近似有更高的粗糙集近似度。因此, 在 IIIS 上建立决策粗糙集模型时采用极大相容类对粗糙集近似算子定义。

2.3 决策粗糙集模型

定义 8 中粗糙集下近似算子较为严格而上近似算子较为宽松, 针对该问题, 本文引入阈值 α, β 分别对集合 X 的下近似与上近似进行限制, 并结合决策粗糙集理论以损失函数的形式对阈值进行表示, 从而建立基于 IIIS 的决策粗糙集模型。

设状态集为 $\{X, \neg X\}$, 代表一个对象非此即彼的两种状态; 设行为集为 $\{a_P, a_B, a_N\}$, 分别表示接受某事物采取的决策行为、待定的决策行为、拒绝某事物采取的决策行为, 即分别确定 $x \in \text{pos}(X)$ 、 $x \in \text{bn}(X)$ 、 $x \in \text{neg}(X)$; 设 $\lambda_{PP}, \lambda_{BP}, \lambda_{NP}$ 分别为 $x \in X$ 时采取行为 a_P, a_B 和 a_N 所产生的风险损失, 相应地, $\lambda_{PN}, \lambda_{BN}, \lambda_{NN}$ 分别为 $x \in \neg X$ 时采取行为 a_P, a_B 和 a_N 所产生的风险损失; 设条件概率为 $P(X \mid M_A^\omega(x)) = \frac{|M_A^\omega(x) \cap X|}{|M_A^\omega(x)|}$ 。则对

$$\begin{aligned} R(a_P \mid M_A^\omega(x)) &= \lambda_{PP}P(X \mid M_A^\omega(x)) + \lambda_{PN}P(\neg X \mid M_A^\omega(x)) \\ R(a_B \mid M_A^\omega(x)) &= \lambda_{BP}P(X \mid M_A^\omega(x)) + \lambda_{BN}P(\neg X \mid M_A^\omega(x)) \\ R(a_N \mid M_A^\omega(x)) &= \lambda_{NP}P(X \mid M_A^\omega(x)) + \lambda_{NN}P(\neg X \mid M_A^\omega(x)) \end{aligned} \quad (10)$$

根据贝叶斯最小风险决策原则可以产生如下决策规则:

(P) 如果 $R(a_P \mid M_A^\omega(x)) \leq R(a_B \mid M_A^\omega(x))$ 并且 $R(a_P \mid M_A^\omega(x)) \leq R(a_N \mid M_A^\omega(x))$, 则 $x \in \text{pos}(X)$ 。

(B) 如果 $R(a_B \mid M_A^\omega(x)) \leq R(a_P \mid M_A^\omega(x))$ 并且 $R(a_B \mid M_A^\omega(x)) \leq R(a_N \mid M_A^\omega(x))$, 则 $x \in \text{bn}(X)$ 。

(N) 如果 $R(a_N \mid M_A^\omega(x)) \leq R(a_P \mid M_A^\omega(x))$ 并且 $R(a_N \mid M_A^\omega(x)) \leq R(a_B \mid M_A^\omega(x))$, 则 $x \in \text{neg}(X)$ 。

考虑到现实中接受正确事物的损失不大于延迟接受正确事物的损失且均小于拒绝正确事物的损失; 同理, 拒绝错误事物的损失不大于延迟拒绝错误事物的损失且均小于拒绝正确事物的损失。因此损失函数满足如下条件:

$$\lambda_{PP} \leq \lambda_{BP} \leq \lambda_{NP}, \lambda_{NN} \leq \lambda_{BN} \leq \lambda_{PN} \quad (11)$$

并且考虑到 $P(X \mid M_A^\omega(x)) + P(\neg X \mid M_A^\omega(x)) = 1$, 则决策规则 (P) ~ (N) 可以整理为

(P) 如果 $P(X \mid M_A^\omega(x)) \geq \alpha$ 并且 $P(X \mid M_A^\omega(x)) \geq \gamma$, 则 $x \in \text{pos}(X)$ 。

(B) 如果 $P(X \mid M_A^\omega(x)) \leq \alpha$ 并且 $P(X \mid M_A^\omega(x)) \geq \beta$, 则 $x \in \text{bn}(X)$ 。

(N) 如果 $P(X \mid M_A^\omega(x)) \leq \beta$ 并且 $P(X \mid M_A^\omega(x)) \leq \gamma$, 则 $x \in \text{neg}(X)$ 。

其中, 阈值 α, β, γ 分别为

$$\begin{aligned} \alpha &= \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{BP} - \lambda_{PP} + \lambda_{PN} - \lambda_{BN}}, \beta = \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{NP} - \lambda_{BP} + \lambda_{BN} - \lambda_{NN}} \\ \gamma &= \frac{\lambda_{PN} - \lambda_{NN}}{\lambda_{NP} - \lambda_{PP} + \lambda_{PN} - \lambda_{NN}} \end{aligned} \quad (12)$$

基于极大相容类的决策粗糙集模型的下近似与上近似分别为

$$\begin{aligned} \underline{\text{apr}}_{A(\alpha, \beta)}^\omega(X) &= \{x \in U \mid P(X \mid M_A^\omega(x)) \geq \alpha\} \\ \overline{\text{apr}}_{A(\alpha, \beta)}^\omega(X) &= \{x \in U \mid P(X \mid M_A^\omega(x)) > \beta\} \end{aligned} \quad (13)$$

由定义 8 与式 (13) 易知, 基于不完备区间值信息系统的决策粗糙集模型具有如下定理。

定理 3 给定不完备区间值信息系统 $K = (U, A, V, F)$, ω 是相似度阈值, $\forall B \subseteq A, X \subseteq U$, 则:

$$\frac{|\underline{\text{apr}}_{B(\alpha, \beta)}^\omega(X)|}{|\overline{\text{apr}}_{B(\alpha, \beta)}^\omega(X)|} \geq \frac{|\underline{\text{APR}}_B^\omega(X)|}{|\overline{\text{APR}}_B^\omega(X)|} \quad (14)$$

定理 4 给定不完备区间值信息系统 $K = (U, A, V, F)$, ω 是相似度阈值, $\forall B \subseteq A, X \subseteq U$, 当 $\alpha = 1$ 时 $\underline{\text{apr}}_{B(\alpha, \beta)}^\omega(X) = \underline{\text{APR}}_B^\omega(X)$, 当 $\beta = 0$ 时 $\overline{\text{apr}}_{B(\alpha, \beta)}^\omega(X) = \overline{\text{APR}}_B^\omega(X)$ 。

由定理 3 可以看出, 基于 IIIS 的决策粗糙集模型较基于极大相容类定义的粗糙集具有更高的粗糙集近似度; 由定理 4 可以看出, 基于 IIIS 的决策粗糙集模型是基于极大相容类粗糙集的延伸与拓展, 更具有一般性。

3 IIIS 下基于区分矩阵的决策粗糙集属性约简

决策粗糙集模型由于分类阈值 α, β 的引入, 使得经典粗糙集模型中决策区域的单调性在决策粗糙集模型中不再成立。为保证对每个对象采取的决策行为不发生变化, 定义不完备区间值决策信息系统下属性约简定义如下。

定义 9 给定不完备区间值决策信息系统 $K = (U, A \cup D, V, F)$, 其中 D 为决策属性, $\pi_D = U/D = \{d_1, d_2, \dots, d_m\}$, ω 是相似度阈值, 则称

$$D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x)) = \{d_i \in \pi_D \mid P(X \mid M_A^\omega(x)) \geq \alpha\} \quad (15)$$

为对象 x 的正域分布函数。

定义 10 给定不完备区间值决策信息系统 $K = (U, A \cup D, V, F)$, 其中 D 为决策属性, ω 是相似度阈值, $\forall B \subseteq A$, 若 $\forall x \in U, D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x)) = D_{\text{POS}(\alpha, \beta)}(M_B^\omega(x))$ 且 $\forall a \in B, D_{\text{POS}(\alpha, \beta)}$

$(M_B^\omega(x)) \neq D_{\text{POS}(\alpha, \beta)}(M_{B-|a|}^\omega(x))$, 则 B 是不完备区间值决策信息系统的一个属性约简。

定义 11 给定不完备区间值决策信息系统 $K = (U, A \cup D, V, F)$, 其中 D 为决策属性, ω 是相似度阈值, 条件属性集 A 下区分矩阵的元素 $m(i, j)$ 定义为

$$m(i, j) = \begin{cases} \emptyset, D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_i)) = D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_j)) \\ \{a_k \mid a_k \in A \wedge s_{ij}^k < \omega\}, \text{ other} \end{cases} \quad (16)$$

根据定义 11 可得到基于极大相容类决策粗糙集的区分矩阵属性约简算法。

算法 1 基于极大相容类决策粗糙集的区分矩阵属性约简算法

输入: 不完备区间值决策信息系统 $K = (U, A \cup D, V, F)$, $\alpha, \beta \in (0, 1)$ 且 $\alpha > \beta, \omega \in (0, 1]$ 。

输出: 属性约简。

1) 根据给定的决策信息系统 $K = (U, A \cup D, V, F)$, 结合相似度阈值 ω 求出所有对象的容差类;

2) 求出所有对象的极大相容类;

3) 结合决策粗糙集分类阈值 α 求出所有对象的正域分布;

4) 构建区分矩阵;

5) 对所有非空区分属性集进行合取, 再将合取转换为析取, 从而求出属性约简。

4 属性约简实例

表 1 为不完备区间值决策信息系统, 其中 $A = \{a_1, a_2, a_3, a_4, a_5\}$ 为条件属性集, D 为决策属性, $U = \{x_1, x_2, \dots, x_{10}\}$ 为论域, 目标集为 $X = \{x \mid x \in U \wedge f_D(x) = 1\}$, 给定信息系统损失函数根据式(12)经计算得 $\alpha = 0.7, \beta = 0.55$, 取相似度阈值 $\omega = 0.7$ 。

表 1 不完备区间值决策信息系统

U	a_1	a_2	a_3	a_4	a_5	D
x_1	[3.27, 4.96]	*	[7.87, 9.21]	[1.54, 3.21]	[8.83, 10.13]	1
x_2	[3.62, 5.08]	[5.62, 7.20]	[8.01, 9.37]	[1.75, 3.46]	[8.97, 10.49]	1
x_3	[4.26, 5.37]	[5.03, 5.91]	*	[1.63, 3.57]	[8.59, 9.97]	2
x_4	[3.00, 4.44]	[5.83, 6.71]	[7.21, 8.21]	[1.54, 3.73]	[7.85, 9.71]	2
x_5	[3.42, 4.78]	[6.09, 6.97]	[8.13, 9.26]	[1.96, 3.87]	*	1
x_6	[3.47, 5.12]	[5.43, 7.01]	[8.05, 9.19]	[1.56, 3.90]	[8.85, 10.13]	2
x_7	[2.97, 4.69]	[5.23, 6.75]	[8.15, 9.14]	*	[8.53, 9.97]	1
x_8	[3.39, 5.48]	[5.31, 7.02]	[8.07, 9.17]	[1.72, 3.58]	[9.02, 10.30]	1
x_9	*	[5.27, 7.15]	[7.98, 9.32]	[1.69, 3.68]	[8.89, 10.25]	2
x_{10}	[3.01, 4.35]	[5.68, 6.56]	[9.01, 9.88]	[1.02, 1.79]	[6.41, 7.52]	2

下面应用算法 1 对该信息系统进行属性约简。具体步骤如下。首先对各对象间属性相似度计算, 以对象 x_1 为例, 根据式(5)可得 x_1 与其他对象之间相似度如表 2 所示。

因此, 当相似度阈值 $\omega = 0.7$ 时, x_1 的容差类为 $\Gamma_A^\omega(x_1) = \{x_1, x_2, x_6, x_8\}$ 。同理可得, $\Gamma_A^\omega(x_2) = \{x_1, x_2, x_6, x_8\}$, $\Gamma_A^\omega(x_3) = \{x_3\}$, $\Gamma_A^\omega(x_4) = \{x_4\}$, $\Gamma_A^\omega(x_5) = \{x_5\}$, $\Gamma_A^\omega(x_6) = \{x_1, x_2, x_6, x_8\}$, $\Gamma_A^\omega(x_7) = \{x_7\}$, $\Gamma_A^\omega(x_8) = \{x_1, x_2, x_6, x_8, x_9\}$, $\Gamma_A^\omega(x_9) = \{x_8, x_9\}$, $\Gamma_A^\omega(x_{10}) = \{x_{10}\}$ 。

由以上计算结果可知, 包含 x_1 的相容类为 $M = \{x_1, x_2, x_6, x_8\}$, 又 $\forall x_m \in U - M$ 必存在 $x_j \in M$ 使得 $(x_j, x_m) \notin \Gamma_A^\omega$, 因此 x_1 的极大相容类为: $M_A^\omega(x_1) = \{x_1, x_2, x_6, x_8\}$ 。

表 2 x_1 与其他对象各属性下属性相似度

	a_1	a_2	a_3	a_4	a_5
x_1	1	1	1	1	1
x_2	0.740	0.728	0.8000	0.760	0.700
x_3	0.333	0.406	0.502	0.778	0.740
x_4	0.597	0.406	0.170	0.763	0.386
x_5	0.805	0.406	0.777	0.537	0.319
x_6	0.805	0.728	0.851	0.700	0.985
x_7	0.714	0.701	0.739	0.580	0.713
x_8	0.710	0.788	0.821	0.730	0.755
x_9	0.673	0.866	0.848	0.710	0.873
x_{10}	0.554	0.406	0.100	0.114	0

同理可得, 其余对象的极大相容类为: $M_A^\omega(x_2) = \{x_1, x_2, x_6, x_8\}$, $M_A^\omega(x_3) = \{x_3\}$, $M_A^\omega(x_4) = \{x_4\}$, $M_A^\omega(x_5) = \{x_5\}$, $M_A^\omega(x_6) = \{x_1, x_2, x_6, x_8\}$, $M_A^\omega(x_7) = \{x_7\}$, $M_A^\omega(x_8) = \{x_1, x_2, x_6, x_8\}$, $M_A^\omega(x_9) = \{x_8, x_9\}$, $M_A^\omega(x_{10}) = \{x_{10}\}$ 。

根据式(15)且分类阈值 $\alpha = 0.7$ 可知, 每个对象的正域分布为: $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_1)) = \{X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_2)) = \{X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_3)) = \{\neg X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_4)) = \{\neg X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_5)) = \{X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_6)) = \{X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_7)) = \{X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_8)) = \{X\}$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_9)) = \emptyset$, $D_{\text{POS}(\alpha, \beta)}(M_A^\omega(x_{10})) = \{\neg X\}$ 。

结合式(16)与以上结论, 构建区分矩阵如表 3 所示。

表 3 不完备区间值决策信息系统区分矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	$\emptyset$									
x_2	$\emptyset$	$\emptyset$								
x_3	$\{a_3\}$	$\{a_2, a_3, a_5\}$	$\emptyset$							
x_4	$\{a_1, a_2, a_3, a_5\}$	$\{a_1, a_2, a_3, a_5\}$	$\emptyset$	$\emptyset$						
x_5	$\emptyset$	$\emptyset$	$\{a_2, a_3\}$	$\{a_1, a_2, a_3, a_5\}$	$\emptyset$					
x_6	$\emptyset$	$\emptyset$	$\{a_2, a_3\}$	$\{a_1, a_2, a_3, a_5\}$	$\emptyset$	$\emptyset$				
x_7	$\emptyset$	$\emptyset$	$\{a_1, a_3, a_4\}$	$\{a_2, a_3, a_5\}$	$\emptyset$	$\emptyset$	$\emptyset$			
x_8	$\emptyset$	$\emptyset$	$\{a_2, a_3, a_5\}$	$\{a_1, a_2, a_3, a_5\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$		
x_9	$\{a_1\}$	$\{a_1\}$	$\{a_2, a_3, a_5\}$	$\{a_1, a_2, a_3, a_5\}$	$\{a_1, a_2, a_5\}$	$\{a_1\}$	$\{a_1, a_4, a_5\}$	$\emptyset$	$\emptyset$	
x_{10}	A	A	$\emptyset$	$\emptyset$	A	A	$\{a_2, a_3, a_4, a_5\}$	A	A	$\emptyset$

对区分矩阵中所有非空属性进行合取, 再将合取转换为析取:

$$L = a_3 \wedge (a_1 \vee a_2 \vee a_3 \vee a_5) \wedge a_1 \wedge (a_1 \vee a_2 \vee a_3 \vee a_4 \vee a_5) \wedge (a_2 \vee a_3 \vee a_5) \wedge (a_2 \vee a_3) \wedge (a_1 \vee a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_5) \wedge (a_1 \vee a_4 \vee a_5) \wedge (a_2 \vee a_3 \vee a_4 \vee a_5) = a_1 \wedge a_3$$

即属性约简集为 $\{a_1, a_3\}$ 。分别基于属性集 $\{a_1, a_3\}$ 、 $\{a_1\}$ 、 $\{a_3\}$ 对信息系统中各对象的正域分布计算, 结果如表 4 所示。

表 4 属性集 $\{a_1, a_3\}$ 、 $\{a_1\}$ 、 $\{a_3\}$ 下各对象正域分布

	$\{a_1, a_3\}$	$\{a_1\}$	$\{a_3\}$		$\{a_1, a_3\}$	$\{a_1\}$	$\{a_3\}$
x_1	$\{X\}$	$\{X\}$	$\{X\}$	x_6	$\{X\}$	$\{X\}$	$\{X\}$
x_2	$\{X\}$	$\{X\}$	$\{X\}$	x_7	$\{X\}$	$\{X\}$	$\{X\}$
x_3	$\{\neg X\}$	$\{\neg X\}$	$\{\neg X\}$	x_8	$\{X\}$	$\{X\}$	$\{X\}$
x_4	$\{\neg X\}$	$\emptyset$	$\{\neg X\}$	x_9	$\emptyset$	$\emptyset$	$\{X\}$
x_5	$\{X\}$	$\{X\}$	$\{X\}$	x_{10}	$\{\neg X\}$	$\emptyset$	$\{\neg X\}$

由表 4 可以看出, 属性集 $\{a_1, a_3\}$ 是保持 (下转第 页)

合疾病名称的识别和抽取更加有效。此外,本文提出的多标签 CRF 算法不仅适用于英文文本,对于中文的医疗文本也同样适用,但是应根据中文医疗文本中疾病名称的复合情况来设计多标签的层数,并针对中文医疗文本的特点选择合适的特征。

4 结束语

本文主要研究了生物医疗文本中疾病名称的识别,使用 multi-label CRF 模型对连续或非连续式和交叉混合式疾病名称进行抽取,并针对医疗文本的特点,尝试了大量的特征,最后通过实验验证了该方法的有效性。

目前此方法仍有一些不足,如 number 标签存在一定的稀疏性,如果数据中出现三个以上实体交叉混合情况时,那么设计的 number 标签将会比较稀疏,这部分有待于加强,需要继续改进多标签。另外,本文选取的特征针对的是英文医疗文本,对于中文医疗文本的疾病名称抽取则需要重新选择合适的特征,同时特征设计方面还可以寻找更多的特征,比如句子级的语义特征,如依存分析等,或者考虑利用外部字典资源。为了更好的识别效果和应用,这些都是本文的下一步工作。

参考文献:

- [1] Hripesak G, Friedman C, Alderson P O, *et al.* Unlocking clinical data from narrative reports: a study of natural language processing [J]. *Annals of Internal Medicine*, 1995, 122(9): 681-688.
- [2] Friedman C, Alderson P O, Austin J H, *et al.* A general natural-language text processor for clinical radiology [J]. *Journal of the American Medical Informatics Association*, 1994, 1(2): 161.
- [3] Mork J G, Bodenreider O, Demner-Fushman D, *et al.* Extracting Rx information from clinical narrative [J]. *Journal of the American Medical Informatics Association*, 2010, 17(5): 536-539.
- [4] Savova G K, Masanz J J, Ogren P V, *et al.* Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications [J]. *Journal of the American Medical Informatics Association*, 2010, 17(5): 507-513.
- [5] Roberts A, Gaizauskas R, Hepple M. Extracting clinical relationships from patient narratives [C]//Proc of the Workshop on Current Trends in Biomedical Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2008: 10-18.
- [6] De Bruijn B, Cherry C, Kiritchenko S, *et al.* Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010 [J]. *Journal of the American Medical Informatics Association*, 2011, 18(5): 557-562.
- [7] Xu Shuo, An Xin, Zhu Lijun, *et al.* A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature [J]. *Journal of Cheminform*, 2014, 7(Suppl 1): S11.
- [8] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proc of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 282-289.
- [9] Jonnagaddala J, Chang Naiwen, Jue T R, *et al.* Recognition and normalization of disease mentions in PubMed abstracts [C]//Proc of the 5th BioCreative Challenge Evaluation Workshop. 2015: 234-239.
- [10] Ratnikov L, Roth D. Design challenges and misconceptions in named entity recognition [C]//Proc of the Thirteenth Conference on Computational Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics, 2009: 147-155.
- [11] Leaman R, Wei C H, Lu Zhiyong. tmChem: a high performance approach for chemical named entity recognition and normalization [J]. *Journal of Cheminformatics*, 2015, 7(Suppl 1): S3.
- [12] Krallinger M, Leitner F, Rabal O, *et al.* CHEMDNER: the drugs and chemical names extraction challenge [J]. *Journal of Cheminformatics*, 2015, 7(Suppl 1): S1.
- [13] Tang Buzhou, Cao Hongxin, Wang Xiaolong, *et al.* Evaluating word representation features in biomedical named entity recognition tasks [J]. *BioMed Research International*, 2014, 2014(2): 6.

(上接第 页)该信息系统正域分布不变的最小属性集,即 $\{a_1, a_3\}$ 是保持信息正域分布不变的一个属性约简。

5 结束语

经典决策粗糙集理论一直在完备的信息系统下,基于等价关系、等价类建立模型,而在不完备区间值信息系统下建立决策粗糙集模型一直是有待讨论的问题,且属性约简一直是决策粗糙集理论中研究的热点。本文首先对不完备区间值信息系统中属性相似度的概念进行改进,进而采用容差关系下极大相容类的概念代替经典决策粗糙集中等价类的概念,实现决策粗糙集理论在不完备区间值信息系统下的拓展并证明该模型具有更高的粗糙集近似度。并且在保持对象正域分布不变的原则下提出基于区分矩阵的属性约简算法,最后将该算法应用于实例。

参考文献:

- [1] Pawlak Z. Rough sets [J]. *International Journal of Computer & Information Sciences*, 1982, 11(5): 341-356.
- [2] 杨霖琳,秦克云.模糊信息系统中一种改进的模糊相容关系[J].小型微型计算机系统,2014,35(9): 2131-2135.
- [3] 郭嗣琮,徐丽,郑爱红.限制容差关系的不完备可变粗糙集[J].辽宁工程技术大学学报:自然科学版,2014(7): 988-991.
- [4] 朱颖东,钟勇.相容关系的改进及其属性约简[J].北京理工大学学报,2010,30(1): 55-59.
- [5] 曾玲,何普彦,付敏.不完备区间值信息系统的粗糙集约简算法[J].南京理工大学学报,2013,37(4): 524-529.
- [6] 赵亮,张欣,薛质.不完备区间值信息系统的安全评估[J].计算机工程,2011,37(11): 146-148.
- [7] 曾玲,付敏,何普彦.双精度容差关系的粗糙集拓展模型及约简[J].江苏大学学报:自然科学版,2014(3): 312-318.
- [8] 刘盾,李天瑞,李华雄.粗糙集理论:基于三支决策视角[J].南京大学学报:自然科学版,2013,49(5): 574-581.
- [9] 张贤勇,苗奇谦.决策粗糙集的一种新分类区域及相关比较分析[J].系统工程理论与实践,2014(12): 3204-3211.
- [10] Liang Decui, Liu Dun, Pedrycz W, *et al.* Triangular fuzzy decision-theoretic rough sets [J]. *International Journal of Approximate Reasoning*, 2013, 54(8): 1087-1106.
- [11] Yao Yiyu, Zhao Yan. Attribute reduction in decision-theoretic rough set models [J]. *Information Sciences*, 2008, 178(17): 3356-3373.
- [12] Jia Xiuyi, Liao Wenhe, Tang Zhenmin, *et al.* Minimum cost attribute reduction in decision-theoretic rough set models [J]. *Information Sciences*, 2013, 219(1): 151-167.