

粒度计算中混合属性约简的权重模糊粗糙集模型

刘洋¹, 周清雷¹, 冯博琴²

(1. 郑州大学信息工程学院, 450001, 郑州; 2. 西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 针对权重粗糙集模型不能有效处理非平衡混合数据的问题, 对权重论域上的各种类型变量进行分析并建立统一的模糊等价关系, 提出混合数据上的权重模糊粗糙集模型, 并利用该模型构造出带权模糊等价空间上的混合属性约简算法. 混合属性约简算法产生的模糊软划分可以克服权重论域上离散硬划分产生的信息损失. 在非平衡混合数据集上进行的实验结果表明, 与基于权重粗糙集的算法相比, 基于权重模糊粗糙集模型的属性约简算法的平均分类精度提高了 11.9%.

关键词: 粗糙集; 混合数据; 模糊等价关系; 权重论域; 属性约简

中图分类号: TP391 **文献标志码:** A **文章编号:** 0253-987X(2011)10-0043-05

A Weighted Fuzzy Rough Sets Model with Hybrid-Attribute Reduction in Granular Computing

LIU Yang¹, ZHOU Qinglei¹, FENG Boqin²

(1. School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: In order to solve the problem that weighted rough sets model lacks a mechanism to deal with mixed and imbalanced data, a unified fuzzy equivalent relationship for analyzing different types of features in weighted domain is established, and a weighted fuzzy rough sets model is proposed to deal with mixed data. Furthermore, a hybrid attribute-reduction algorithm is constructed based on the weighted fuzzy rough sets model. Compared with the classical crisp partition, the hybrid algorithm can avoid information loss through fuzzy soft partition generated by the model. Experimental results on imbalanced and mixed data sets show that the proposed weighted fuzzy rough sets model can not only select fewer features than weighted rough sets model, but also improve the average classification performance of the reduced attribute set on learning methods by 11.9%.

Keywords: rough set; mixed data; fuzzy equivalent relationship; weighted domain; attribute reduction

由 Z. Pawlak 教授^[1]提出的粗糙集理论被广泛应用于知识依赖性分析、属性子集选择等问题^[2]. 遗憾的是, 作为一种有效的粒度计算模型, Pawlak 粗糙集定义在经典论域的基础上, 因此缺乏一种能够有效地结合论域上的先验信息和主观知识的机制^[3], 对于现实应用中广泛存在的先验性知识不能

直接处理. 此外, 在工程、医疗和金融等领域, 混合型变量无处不在. 研究人员在引入粗糙集等机器学习方法来处理该类数据时, 往往采用离散化算法把数值型变量和模糊型变量转化为符号型变量^[4], 然而这一转换不可避免地带来了信息损失, 学习算法的性能在很大程度上取决于离散化的效果^[5].

为了解决这一问题,人们引入了模糊粗糙集模型^[6]和邻域粗糙集模型^[7].不同的模型基于不同的粒度度量标准和逼近定义,但它们都属于粒度计算的研究范畴^[8].由于这些方法仅适用于普通论域上的粒度度量,因此都缺乏处理主观知识和先验信息的能力.文献[9-11]研究了结合权重信息处理粗糙集属性约简的问题,然而这些方法仅适用于符号属性的情形.文献[12]基于带权论域上的模糊等价关系,讨论了带权模糊近似空间,并给出有效的规则获取方法.但是,带权模糊近似空间的属性约简问题还未查到有文献进行分析和研究.

本文利用模糊等价关系的信息熵度量标准,提出一种基于权重模糊粗糙集模型的混合属性约简算法.该算法直观、易于理解,能够同时处理符号型、数值型和模糊型变量,而无需对其进行离散化处理和去模糊化处理.因此,与经典的粗糙集算法相比,该算法结合了论域上的权重信息,而且省去了混合型属性的预处理过程.由于该模型可以结合先验知识并直接分析混合数据,因而拓展了经典粗糙集理论的应用范围.

1 各类型变量空间的统一模糊粒度化表示

粒度表示是粗糙集理论和粒度计算的基本问题,Pawlak 粗糙集模型建立在离散空间的不可辨识等价关系之上,等价关系对论域的划分形成了论域空间的粒度化表示.然而,对于数值型、模糊型变量空间,对象的取值可以是连续的、模糊的,采用等价关系将导致对个别属性的过拟合.本文所讨论的模型建立在模糊等价关系的粒度化表示基础上.

定义 1 给定一个非空对象集合 $U = \{x_1, \dots, x_m\}$, $\tilde{R} = (r_{ij})_{m \times m}$ 是 U 上的一个模糊二元关系,其中 $r_{ij} \in [0, 1]$ 为对象 x_i 和 x_j 之间关系的度量.

一般来讲,如果对于所有的 $x_i, x_j, x_k \in U, \tilde{R}$ 满足自反性 $r_{ii} = 1$ 、对称性 $r_{ij} = r_{ji}$ 和传递性 $r_{ik} = \bigwedge_j (r_{ij} \wedge r_{jk})$,则称 $\tilde{R}$ 是一个模糊等价关系.

定义 2 由模糊等价关系 $\tilde{R}$ 产生的模糊等价类分为 $U/\tilde{R} = \{[x_i]_{\tilde{R}}\}$,其中 $[x_i]_{\tilde{R}} = \int r_{ij}/x_j$ 为包含对象 x_i 的模糊等价类, $\int$ 表示模糊集合元素的并.

由定义 2 可以看出,模糊等价关系是经典等价

关系的扩展.对于符号型属性,其产生的模糊等价关系就是 Pawlak 粗糙集模型描述的经典等价关系.对于数值型属性,通常需要先对连续特征进行归一化操作.L. Zadeh 教授^[13]提出了一种基于相似性度量函数的模糊相似二元关系 $\text{sim}(i, j) = g(|x_i(t) - x_j(t)|)$,这里函数 g 满足^[13]

$$g(0) = 1; g(\infty) = 0; g(\cdot) \in [0, 1] \quad (1)$$

$$\text{sim}(i, j) = \text{sim}(j, i); \text{sim}(i, i) = 1 \quad (2)$$

函数 g 有多种选择,本文采用如下定义

$$g(d) = \begin{cases} 1 - d/\delta, & d < \delta \\ 0, & \text{其他} \end{cases} \quad (3)$$

式中: $\delta \in [0, 0.5]$,本文设置 $\delta = 0.25$.

对于模糊型属性,有许多可供选择的模糊相似性度量标准,例如 Hamming 相似性度量函数或者最大-最小相似度量函数,其时间复杂度为 $O(n^2)$ ^[14].

2 权重模糊粗糙集模型

文献[12]基于带权论域上的模糊等价关系,研究了带权模糊近似空间的上、下近似,基于带权模糊近似空间提出了有效的规则获取方法.本节将在此基础上,对带权论域上的属性依赖问题进行深入的讨论研究.

2.1 权重信息系统

定义 3 权重信息系统为一个五元组 $\langle U, A, V, f, \omega \rangle$,其中 U 是非空对象集合, $A = \{a_1, \dots, a_n\}$ 为非空属性集合, $f: U \times A \rightarrow V$ 为一个信息函数, V 为属性值域的集合, $\omega: U \rightarrow R$ 为论域 U 上的权重分配函数.通常,二元组 $\langle U, \omega \rangle$ 被称为权重信息系统的带权论域空间,若 $A = C \cup D$,其中 C, D 分别为条件和决策属性集合,则权重信息系统也可称为权重决策信息系统.

定义 4 给定一个带权论域 $\langle U, \omega \rangle$ 上的一个模糊集合 $\tilde{X}, \tilde{Y}$ 的带权度数定义为

$$|\tilde{X}|_{\omega} = \sum_i \frac{\omega_i}{\sum_j \omega_j} \tilde{X}(i) \quad (4)$$

定理 1 给定一个带权论域 $\langle U, \omega \rangle$ 上的 2 个模糊集合 $\tilde{X}, \tilde{Y}$,如果 $\tilde{X} \subseteq \tilde{Y}$,则以下公式成立

$$0 \leq |\tilde{X}|_{\omega} \leq 1 \quad (5)$$

$$|\tilde{X}|_{\omega} \leq |\tilde{Y}|_{\omega} \quad (6)$$

$$|\tilde{X} \cap \tilde{Y}|_{\omega} = |\tilde{X}|_{\omega} \quad (7)$$

$$|\tilde{X} \cup \tilde{Y}|_{\omega} = |\tilde{Y}|_{\omega} \quad (8)$$

文献[12]研究了带权模糊近似空间的上、下近似,本文不再赘述.下面对带权论域上的属性依赖性度量进行讨论.

定义 5 给定一个权重决策信息系统 $\langle U, A, V, f, w \rangle$, $B \subseteq C$, L_{pos} 为属性集的正域, B 相对于决策属性 D 的带权依赖度定义为

$$\gamma_B(D, w) = |L_{\text{pos}}(D)|_w \tag{9}$$

根据定义可知, $\gamma \in [0, 1]$. 当 $\gamma = 1$ 时, 决策属性集合 D 完全依赖于 B ; 当 $0 < \gamma < 1$ 时, 决策属性集合 D 部分依赖于 B ; 否则, $\gamma = 0$ 时, 决策属性集合 D 不依赖于 B .

2.2 属性依赖程度的信息熵度量

信息熵已经广泛应用于粗糙集理论的属性重要性度量中, 但是对于带权论域上模糊等价关系的信息熵度量方式, 需要给出新的定义来刻画属性在带权论域空间上的分辨能力.

定义 6 给定一个带权论域 $\langle U, w \rangle$, 对于带权论域上的任意一个模糊等价关系 $\tilde{R}$, 其信息量函数定义为

$$H(\tilde{R}, w) = - \sum_i \frac{w_i}{\sum_j w_j} \lg |[x_i]_{\tilde{R}}|_w \tag{10}$$

定理 2 给定权重信息系统, 其中 $B, E \subseteq C$, 由 B, E 产生的模糊等价关系分别为 $\tilde{R}, \tilde{S}$, 则集合 B, E 的联合信息熵为

$$H(B \cup E, w) = - \sum_i \frac{w_i}{\sum_j w_j} \lg |[x_i]_{\tilde{R}} \cap [x_i]_{\tilde{S}}|_w \tag{11}$$

证明 由定义可知 $H(B \cup E, w) = H(\tilde{R} \cap \tilde{S}, w) = - \sum_i \frac{w_i}{\sum_j w_j} \lg |[x_i]_{\tilde{R} \cap \tilde{S}}|_w$, 因为 $[x_i]_{\tilde{R}} \cap [x_i]_{\tilde{S}} = [x_i]_{\tilde{R} \cap \tilde{S}}$, 故定理 2 得证.

定义 7 给定权重信息系统, 其中 $B, E \subseteq C$, 由 B, E 产生的模糊等价关系分别为 $\tilde{R}, \tilde{S}$, 则集合 B, E 的条件信息熵为

$$H(E | B, w) = - \sum_i \frac{w_i}{\sum_j w_j} \lg \frac{|[x_i]_{\tilde{R}} \cap [x_i]_{\tilde{S}}|_w}{|[x_i]_{\tilde{R}}|_w} \tag{12}$$

2.3 基于权重模糊粗糙集的混合属性约简算法

发现权重决策信息系统的全部约简需要测试 $2^n - 1$ 个属性子集, 以检验它们是否满足约简的条件, 其中 n 是条件属性的数量. 这对于属性子集较多

的决策信息系统而言, 计算量会大到不能容忍. 本节将基于属性子集的信息增益, 构造一个前向贪心约简算法. 信息增益函数描述了条件属性对分类的贡献, 因此可以作为属性集合重要性度量的评价指标.

定义 8 给定一个权重决策信息系统, 其中 $B \subseteq C$, $a \in C \setminus B$, 属性 a 相对于属性集 B 的相对重要度定义为

$$G_w(a, B, D) = H(B | D, w) - H(B \cup \{a\} | D, w) \tag{13}$$

定理 3 给定一个权重决策信息系统, 其中 $B \subseteq C$, $a \in C \setminus B$, 则 $G_w(a, B, D) \geq 0$.

证明 只需证明 $H(B | D, w) \geq H(B \cup \{a\} | D, w)$. 因为 $[x_i]_{\tilde{R}_B} \cap [x_i]_{\tilde{R}_D} \supseteq [x_i]_{\tilde{R}_{B \cup \{a\}}} \cap [x_i]_{\tilde{R}_D}$, 由定理 1 有 $|[x_i]_{\tilde{R}_B} \cap [x_i]_{\tilde{R}_D}|_w \geq |[x_i]_{\tilde{R}_{B \cup \{a\}}} \cap [x_i]_{\tilde{R}_D}|_w$, 因此有 $H(B | D, w) \geq H(B \cup \{a\} | D, w)$. 定理 3 得证.

图 1 给出本文基于权重模糊粗糙集模型的混合属性约简算法 (WFRS-AR), 它能够确保重要的属性首先被加入到备选约简集合中, 从而不损失重要的属性特征.

输入: 权重决策信息系统 $\langle U, A, V, f, w \rangle$, 阈值参数.
输出: 一个属性约简集合 E .
步骤1: 对于所有的 $a \in C$, 计算 $H(a | D, w)$;
步骤2: $a = \arg \max_{a \in C} H(a | D, w)$, $E = \{a\}$;
步骤3: 对于所有的 $a \in C \setminus E$, 计算 $G_w(a, E, D)$;
步骤4: $a = \arg \max_{a \in C \setminus E} G_w(a, E, D)$;
步骤5: 如果 $G_w(a, E, D) < \varepsilon$, 返回 E ; 否则 $E = E \cup \{a\}$, 转步骤3.

图 1 基于权重模糊粗糙集的混合属性约简算法

3 实验分析

为验证基于权重模糊粗糙集约简算法的有效性, 本文从 UCI 机器学习数据库^[15]中挑选了 4 组数据, 描述如表 1 所示. 可以看出, 这 4 个分类问题都是混合型数据, 并且各类之间的分布是不平衡的.

定义 9 给定 m_- 、 m_+ 分别为少数类、多数类的用例

表 1 数据集描述					
数据集	缩写	实例数	符号属性个数	连续属性个数	两类分布比例
StatLog heart disease	hea	270	6	7	120 : 150
Horse Colic database	hor	368	20	7	124 : 244
Australian credit approval	aca	690	8	6	307 : 383
Hypothyroid disease	hyp	3 163	18	7	151 : 3 012

数, W_- 、 W_+ 分别为少数类和多数类上用例的权重, 少数类权重总和所占的比例定义为

$$P_- = \frac{m_- W_-}{m_- W_- + m_+ W_+}$$

(14)

由定义可知, 随着 P_- 的增加, 少数类用例的权重和也相应增加. 当 $P_- = 0.5$ 时, 逆类概率权重分配给了每一个类, 此时数据集通过权值分配达到了各类之间的平衡效果. 为比较本文算法 (WFRS-AR 算法) 和基于权重粗糙集的算法 (WRS-AR 算法)^[11] 在非平衡混合数据集上选择特征的分类能力, 将 2 种约简算法分别作为当前流行的基于权重的 RBF-SVM 分类学习算法^[16] 的前端属性约简工具. 符号和数值属性的粒度表示采用了本文第 2 节的方法. 本文算法的 ϵ 取值 0.012 5. 实验中, 固定少数类 W_- 的权值, 选择 W_+ 使 P_- 依次为 0.05, 0.1,

0.15, ..., 0.9, 0.95, 记录每次实验中学习算法在约简数据集上的真正率、假正率对, 利用 3 次样条插值积分法计算分类算法的平均分类精度 AUC (area under ROC curves), 评价 2 种约简算法选择特征的质量.

表 2 给出了 2 种属性约简算法在数据集上选择的特征集结果, 以及作为学习器的前端约简工具所取得的平均分类精度. 由表 2 可以发现, 这 2 种算法都能有效地降低特征数量, 相对而言, WRS-AR 算法得到的特征数量较少, 但平均分类性能也较低. 由于 WFRS-AR 算法不受离散化预处理的影响, 通过为不同类别实例分配合适的权重, 在保留了相对较少的特征数的同时, 约简属性集在分类算法上的平均分类精度提高了 11.9%.

表 2 2 种算法的实验结果

数据集	特征数			平均分类精度/%		
	原始数据	WFRS-AR 算法	WRS-AR 算法	原始数据	WFRS-AR 算法	WRS-AR 算法
hea	13	9	6	80.8	86.4	83.3
hor	27	11	10	46.9	69.3	48.2
aca	14	7	5	92.5	87.9	85.2
hyp	25	12	8	73.8	77.6	70.4

4 结论与展望

本文基于模糊等价关系的信息熵概念提出一种分析混合数据的权重模糊粗糙集模型. 该模型以带权论域空间中的模糊等价关系形成论域的粒化, 从而将先验知识引入混合数据的粒度计算模型. 本文进一步给出了带权论域上模糊等价关系的信息量度量方法, 基于信息量增益展示了本文模型在混合属性约简选择中的应用. 实验分析表明, 通过对用例进行权重分配, 权重模糊粗糙集不仅可以选择很少量的特征, 而且可以改善分类器在非平衡数据集上的平均分类性能, 验证了本文算法的有效性. 后续的工作将研究本文模型在代价敏感学习问题中的应用, 此外改进本文模型的海量数据快速约简算法也是研究方向之一.

参考文献:

[1] PAWLAK Z, SKOWRON A. Rudiments of rough sets [J]. Information Sciences, 2007, 177(1): 3-27.
[2] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约

简[J]. 计算机学报, 2002, 25(7): 759-766.
WANG Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy [J]. Chinese Journal of Computers, 2002, 25(7): 759-766.
[3] TING K. An instance-weighting method to induce cost-sensitive trees [J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 659-665.
[4] 谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报, 2005, 28(9): 1570-1574.
XIE Hong, CHENG Haozhong, NIU Dongxiao. Discretization of continuous attributes in rough set theory based on information entropy [J]. Chinese Journal of Computers, 2005, 28(9): 1570-1574.
[5] HU Qinghua, Xie Zongxia, YU Daren. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation [J]. Pattern Recognition, 2007, 40(12): 3509-3521.
[6] DUBOIS D, PRADE H. Rough fuzzy sets and fuzzy rough sets [J]. International Journal of General Systems, 1990, 17(2): 191-209.

- [7] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
HU Qinghua, YU Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation [J]. Chinese Journal of Software, 2008, 19(3): 640-649.
- [8] HU Qinghua, LIU Jinfu, YU Daren. Mixed feature selection based on granulation and approximation [J]. Knowledge-Based Systems, 2008, 21(4): 294-304.
- [9] XU Changzhi, MIN Fan. Weighted reduction for decision tables [C]//Proceedings of 3rd International Conference on Fuzzy Systems and Knowledge Discovery. Berlin, Germany: Springer-Verlag, 2006: 246-255.
- [10] HU Qinghua, YU Daren, XIE Zongxia, et al. Fuzzy probabilistic approximation spaces and their information measures [J]. IEEE Transactions on Fuzzy Systems, 2006, 14(2): 191-201.
- [11] LIU Jinfu, HU Qinghua, YU Daren. A weighted rough set based method developed for class imbalance learning [J]. Information Sciences, 2008, 178(4): 1235-1256.
- [12] LIU Yang, FENG Boqin, BAI Guohua. Compact rule learner on weighted fuzzy approximation spaces for class imbalanced and hybrid data [C]//Proceedings of 6th International Conference on Rough Sets and Current Trends in Computing. Berlin, Germany: Springer-Verlag, 2008: 262-271.
- [13] ZADEH L. Fuzzy sets [J]. Information and Control, 1965, 8(3): 338-353.
- [14] LEE H. An optimal algorithm for computing the max-min transitive closure of a fuzzy similarity matrix [J]. Fuzzy Sets and Systems, 2001, 123(1): 129-136.
- [15] FRANK A, ASUNCION A. UCI machine learning repository [DB/OL]. [2010-12-22]. <http://archive.ics.uci.edu/ml>.
- [16] BREFELD U, GEIBEL P, WYSOTZKI F. Support vector machines with example dependent costs [C]//

Proceedings of the European Conference on Machine Learning. Berlin, Germany, Springer-Verlag, 2003: 23-34.

[本刊相关文献链接]

容粒粒子滤波算法及其应用. 西安交通大学学报, 2011, 45(8): 13-17.

逆向工程中一种鲁棒的平移对称提取算法. 西安交通大学学报, 2011, 45(7): 70-75.

一种鲁棒的子空间聚类算法. 西安交通大学学报, 2011, 45(6): 13-19.

引入模糊推理与强跟踪滤波技术的变结构多模型估计. 西安交通大学学报, 2011, 45(6): 27-33.

一种采用偏最小二乘回归的情绪调节策略预测模型. 西安交通大学学报, 2011, 45(6): 46-49.

采用任务剖面的复杂可修系统保障性仿真与评价技术研究. 西安交通大学学报, 2011, 45(4): 53-59.

三角 Bézier 曲面粗加工刀轨生成算法. 西安交通大学学报, 2011, 45(3): 70-74.

广义证据理论中的基本概率指派生成方法. 西安交通大学学报, 2011, 45(2): 34-38.

面向重用的三维 CAD 模型检索算法. 西安交通大学学报, 2011, 45(1): 74-78.

一种准线性光束平差方法. 西安交通大学学报, 2010, 44(12): 1-4.

广义证据理论的基本框架. 西安交通大学学报, 2010, 44(12): 119-124.

利用特征子空间评价与多分类器融合的高光谱图像分类. 西安交通大学学报, 2010, 44(8): 20-24.

高斯衍生粒子滤波器. 西安交通大学学报, 2010, 44(6): 72-77.

一种多源遥感图像分割的融合新策略. 西安交通大学学报, 2010, 44(6): 88-92.

图像分割的谱聚类集成算法. 西安交通大学学报, 2010, 44(6): 93-98.

应用变量因果序分析的符号有向图建模方法. 西安交通大学学报, 2010, 44(5): 85-90.

(编辑 刘杨 武红江)

《科研论文写作新解》出版

本刊讯 《西安交通大学学报》副主编赵大良编著的《科研论文写作新解》已于 2011 年 9 月由西安交通大学出版社出版. 该书以主编和审稿人的视角诠释了科研论文的写作, 作为科研论文写作教材的辅助读物, 解决了作者与编者之间的信息不对称问题, 适合于具有一定写作基础或投稿经历的读者阅读. 该书尽管是为论文作者而写, 但对于普通期刊编辑认识学术出版的本质、增强与作者沟通的能力、提高办刊水平也具有很好的参考价值.