文章编号:1000－2367(2018)03－0112－07 **DOI**:10.16366/j.cnki.1000－2367.2018.03.019

类别近似质量约束下的属性约简方法研究

李智远[1], 杨习贝[2a], 陈向坚[2a], 王平心[2b]

(1.江苏师范大学 科文学院, 江苏 徐州 221116; 2.江苏科技大学 a.计算机学院;b.理学院, 江苏 镇江 212003)

摘 要:利用近似质量作为度量标准,借助启发式算法求解约简,其本质是根据近似质量的变化情况来找出冗余属性,但这一方法其并未考虑每一个决策类别所对应的下近似集合在约简前后的变化程度.鉴于此,提出了一种基于类别近似质量的属性约简策略,其目标是使得每一个类别的近似质量都满足约简的约束条件.借助邻域粗糙集模型,在 UCI 数据集上将传统约简策略与类别近似质量约简策略进行了对比分析,实验结果不仅验证了类别近似质量约简策略的有效性,而且表明这种策略依然能够满足传统约简的约束条件.

关键词:属性约简;类别近似质量;启发式算法;粗糙集

中图分类号:TP18 **文献标志码**:A

粗糙集[1]是波兰学者 Pawlak 提出的一种用以刻画不确定性的建模方法.经典 Pawlak 粗糙集模型是建立在等价关系基础上的,仅能用于处理符号型数据,但对于现实中广泛存在的连续型数据却缺乏应对策略.因此,借鉴距离的概念,从构建邻域的角度出发,文献[2]提出了邻域粗糙集方法,该方法不仅可以用于直接处理连续型数据,而且极大地拓展了粗糙集方法的应用范畴[3－9].

无论是在经典粗糙集还是邻域粗糙集的研究进程中,属性约简[10－16]问题一直占据着核心地位.所谓属性约简,就是依据粗糙集或依据与学习相关的某种度量指标设置一个约束条件,进而删除数据中的冗余属性,其目的是简化后续问题处理、加速问题求解或提升学习模型的泛化性能.目前,在粗糙集理论中常用的度量指标有近似质量[17]、近似分布[18]、条件熵[19]等,约束条件有保持度量不变或使得度量指标的变化在给定阈值范围内.相比于机器学习领域中的其他特征选择方法,属性约简提供了不同度量标准意义下属性的语义解释,这是属性约简概念的重要意义.

在各种不同的度量指标中,近似质量是以粗糙集的视角来衡量数据中确定性程度的一种技术手段.从近似质量的角度出发,关于粗糙集约简的研究是基于数据中所有决策类所生成下近似并集的变化程度来考虑的.当近似质量在属性变化的情形下呈单调变化趋势时,近似质量约简可以是相较于原始属性集合来说,找到一个使得近似质量不发生变化的最小属性子集;而当近似质量在属性变化的情形下不呈单调变化趋势时,近似质量约简可以是相较于原始属性集合来说,找到一个使得近似质量能够提升的最小属性子集.然而遗憾的是这种约简策略并未考虑每个决策类别的下近似集在约简前后的变化,所以它并一定不能够保证每一个决策类别的下近似集还能在约简后依然能够保持或扩大.例如,面向不平衡数据,在约简之后,虽然数据中的近似质量能够满足给定的约束条件,但是有可能小类数据的下近似集会变得很小甚至为空集,这是一种典型的小类淹没现象,其背后可能存在的原因是因为数据分布的不平衡,所以致使大类类别近似集的约束条件要比小类类别近似集的约束条件相对容易满足.为解决这一问题,在文献[20－21]工作的基础上,利用单个决

收稿日期:2017-09-01;**修回日期**:2017-11-23.

基金项目:国家自然科学基金(61572242;61502211;61503160);中国博士后科学基金(2014M550293);江苏省青蓝工程人才项目.

作者简介:李智远(1978－),男,江苏徐州人,江苏师范大学讲师,主要研究方向为智能信息处理,E-mail:zhiyuan1111@163.com.

通信作者:杨习贝(1980－),男,江苏镇江人,江苏科技大学副教授,博士(后),主要研究方向为粗糙集理论、粒计算、机器学习,E-mail: zhenjiangyangxibei@163.com.

策类别,重新定义了类别近似质量以及相应的属性约简概念.

本文主要内容安排如下:第 1 节简要介绍粗糙集模型;第 2 节提出了类别近似质量约简的定义并设计了启发式算法用以求解类别近似质量约简;第 3 节将类别近似质量约简与传统的近似质量约简进行了实验对比分析;第 4 节总结全文.

1　粗糙集理论

1.1　经典粗糙集

在粗糙集理论中,一个决策系统可以表示为二元组 $DS=<U,AT\cup\{d\}>$:$U=\{x_1,x_2,\cdots,x_n\}$ 是一个非空有限的样本集合,即论域;AT 是所有条件属性集合;d 是决策属性且 $AT\cap\{d\}=\varnothing$. $\forall x_i\in U$,$d(x_i)$ 表示样本 x_i 的类别标记. $\forall A\subseteq AT$,可定义属性集合 A 上的不可分辨关系(等价关系)为

$$\mathrm{IND}(A)=\{(x_i,x_j)\in U^2:\forall a\in A,a(x_i)=a(x_j)\}. \tag{1}$$

类似地,$\mathrm{IND}(\{d\})$ 也是根据决策属性 d 所得到的一个等价关系,因此 $U/\mathrm{IND}(\{d\})=\{X_1,X_2,\cdots,X_q\}$ 可被视作是根据决策属性 d 所得到 U 上的划分,该划分中的每一个等价类表示了一个决策类别范畴,为简化问题描述,如无特殊说明,文中决策类别的下标就表示该决策类所对应的类别标记.

定义 1　给定一个决策系统 DS,$\forall X_p U/\mathrm{IND}(\{d\})$,$\forall A\subseteq AT$,$X_p$ 关于属性集合 A 的下近似集与上近似集分别定义为

$$\underline{X_p}_A=\{x_i\in U:[x_i]_A\subseteq X_p\}, \tag{2}$$

$$\overline{X}_{p_A}=\{x_i\in U:[x_i]_A\cap X_p\neq\varnothing\}. \tag{3}$$

在定义 1 中,$[x_i]_A=\{x_j\in U:(x_i,x_j)\in\mathrm{IND}(A)\}$ 表示 U 中所有与样本 x_i 具有不可分辨关系 $\mathrm{IND}(A)$ 的样本的合集,即 x_i 的等价类.

1.2　邻域粗糙集

类似于经典粗糙集方法,邻域粗糙集的处理对象依然可以表示为决策系统.给定决策系统 $DS=<U,AT\cup\{d\}>$,邻域是建立在某一种度量标准上,通过给定半径考察样本的邻居.不妨假设 $\boldsymbol{M}=(r_{ij})_{n\times n}$ 为论域上的相似度矩阵,r_{ij} 表示样本 x_i 与 x_j 之间的距离度量,给定参数 $\delta\in[0,1]$,$\forall x_i\in U$,x_i 的邻域半径为:

$$\mathrm{Int}(x_i)=\min_{1\le j\le n,j\neq i}r_{ij}+\delta\times(\max_{1\le j\le n,j\neq i}r_{ij}-\min_{1\le j\le n,j\neq i}r_{ij}), \tag{4}$$

其中 $\min\limits_{1\le j\le n,j\neq i}r_{ij}$ 表示 $U-\{x_i\}$ 中的样本与样本 x_i 的距离的最小值,$\max\limits_{1\le j\le n,j\neq i}r_{ij}$ 表示 $U-\{x_i\}$ 中的样本与样本 x_i 的距离的最大值.采用邻域区间的方式考察样本的邻居可以避免因邻域半径过小而产生邻域为空集的情形.借助邻域区间,$\forall x_i\in U$,其邻域为:

$$\delta(x_i)=\{x_j\in U:x_j\neq x_i,r_{ij}\le\mathrm{Int}(x_i)\}. \tag{5}$$

$\delta(x_i)$ 记录了所有与 x_i 之间的距离小于给定邻域区间的样本,可以认为是所有与 x_i 在邻域区间内相似的样本集合.从粒计算的视角来看,$\delta(x_i)$ 是一种信息粒的表现形式,因而也可称为邻域信息粒.

定义 2[2]　给定一个决策系统 DS,根据 d 可以得到所有决策类的合集形如 $\{X_1,X_2,\cdots,X_q\}$,$\forall A\subseteq AT$,d 关于 A 的下近似集和上近似集定义为:

$$\underline{d}_A^N=\bigcup_{p=1}^{q}\underline{X_p}_A^N, \tag{6}$$

$$\overline{d}_A^N=\bigcup_{p=1}^{q}\overline{X_p}_A^N, \tag{7}$$

其中对于任一决策类别 $X_p\in U/\mathrm{IND}(\{d\})$,

$$\underline{X_p}_A^N=\{x_i\in U\mid\delta_A(x_i)\subseteq X_p\}, \tag{8}$$

$$\overline{X_p}_A^N=\{x_i\in U\mid\delta_A(x_i)\cap X_p\neq\varnothing\}. \tag{9}$$

定义 3　给定一个决策系统 DS,$\forall A\subseteq AT$,d 相对于对 A 的近似质量为:

$$\gamma(A,d)=\frac{|\underline{d}_A^N|}{|U|}. \tag{10}$$

其中 $|X|$ 表示集合 X 的基数.

显然 $0 \leq \gamma(A,d) \leq 1$ 成立.$\gamma(A,d)$ 表示根据条件属性集合 A,那些确定属于某一决策类别的样本占总体样本的比例.若 $\underline{d}_A^N$ 越大,则近似质量越高,样本空间的不确定性程度越低.

2 属性约简

2.1 近似质量约简

属性约简是粗糙集理论研究的重要内容.本节中采用近似质量作为度量标准删除冗余属性.由文献[2]可知,在邻域粗糙集上的近似质量满足单调性,即随着条件属性的增加,利用邻域粗糙集所得到的近似质量不断增大.根据大量的实验分析,发现在属性约简中,要求约简前后近似质量完全相等的条件往往过于苛刻,并不利于冗余属性的发现与删除,因此可以采用阈值的方式来控制近似质量约简的约束条件,定义为根据约简后所得到的近似质量与原始数据中的近似质量的差距不超过 $1-\varepsilon(\varepsilon \in [0,1])$.

定义 4 给定决策系统 $DS=<U,AT \cup \{d\}>$,$\forall A \subseteq AT$,A 被称为一个近似质量约简当且仅当

(1)$\gamma(A,d) \geq (1-\varepsilon) \cdot \gamma(AT,d)$;

(2)$\forall \subset A, \gamma(B,d) < (1-\varepsilon) \cdot \gamma(AT,d)$.

定义 4 所示的约简实际上是一个能够使得近似质量的误差在阈值允许范围内的最小属性子集.

2.2 类别近似质量约简

定义 4 中考虑的是在决策系统中,由所有决策类所生成下近似而得到的近似质量,并没有充分考虑每一个决策类别的下近似在约简前后的变化程度.

在属性约简后,如果仅仅保持了近似质量的误差在阈值允许范围内,那么并不一定能够保证每个决策类别的下近似在约简前后都不发生较大的变化,所以一种合理的考虑应该是将每一决策类别的下近似加以单独分析,以期使得每个决策类别上的下近似在约简前后都能够不发生较大的变化.据此以下将给出类别近似质量的公式,用以量化地反映每一个决策类下近似集的大小.给定决策系统 $DS=<U,AT \cup \{d\}>$,根据 d 可以得到所有决策类的合集形如 $U/\mathrm{IND}(\{d\})$,$\forall A \subseteq AT$,$\forall X_p \in U/\mathrm{IND}(\{d\})$,类别 X_p 相对于对 A 的近似质量表示为

$$\gamma(A,X_p)=\frac{|\underline{X_p}_A^N|}{|X_p|}. \tag{11}$$

(11)式描述的是在决策系统中第 p 类样本的近似质量,这是一种基于类别标记的近似质量计算方法.借助(11)式,可以定义如下所示的类别近似质量约简.

定义 5 给定决策系统 $DS=<U,AT \cup D>$,$\forall A \subseteq AT$,针对第 p 个类别标记,A 被称为一个类别近似质量约简当且仅当

(1)$\gamma(A,X_p) \geq (1-\varepsilon) \cdot \gamma(AT,X_p)$;

(2)$\forall B \subset A, \gamma(B,X_p) < (1-\varepsilon) \cdot \gamma(AT,X_p)$.

定义 5 所示属性约简的定义,其目的不是为了能够保持原始决策系统中近似质量的误差在阈值范围内,而是为了保持第 p 类样本的近似质量的误差能够在阈值范围内的最小属性子集.

2.3 启发式算法

给定决策系统 $DS=<U,AT \cup \{d\}>$,$\forall A \subseteq AT$,$\forall aA \in TA-A$,定义属性重要度 $\mathrm{Sig}(a,A,D)$ 用以表示将属性 a 加入到条件属性集合 A 中后近似质量的变化情况,即:

$$\mathrm{Sig}(a,A,D)=\gamma(A \cup \{a\},D)-\gamma(A,D). \tag{12}$$

(12)式所示的属性重要度是根据决策系统中近似质量变化而设计的.类似地,对于第 p 个类别标记,给出如下所示的属性重要度公式.

$$\mathrm{Sig}(a,A,X_p)=\gamma(A \cup \{a\},X_p)-\gamma(A,X_p). \tag{13}$$

基于属性重要度指标,可以构造启发式属性约简算法.该算法以空集为起点,每次计算剩余的每一个属性的属性重要度,从中选择最大的属性重要度所对应的属性,并将其加入约简集合中,直到利用当前约简集

合所求得的近似质量满足约简终止条件.具体步骤如算法 1、算法 2,分别用于求解定义 4 和定义 5 所示的约简.

算法 1　求解近似质量约简.

输入　决策系统 DS,邻域半径参数 δ.

输出　一个约简 red.

步骤 1　由(10)式计算 $\gamma(AT,d)$;

步骤 2　令 $red \leftarrow \varnothing$,$\gamma(red,d)=0$;

步骤 3　若 $\gamma(red,d)<(1-\varepsilon)\cdot\gamma(AT,d)$,则执行以下循环,否则转步骤 4;

(1) $\forall a \in AT-red$,计算属性 a 的重要度 $\mathrm{Sig}(a,red,d)$;

(2) 选择属性 b,满足 $\mathrm{Sig}(b,red,d)=\max\{\mathrm{Sig}(a,red,d),\forall a \in AT-red\}$,$red=red \cup \{b\}$;

(3) 由(10)式计算 $\gamma(red,d)$;

步骤 4　输出 red.

算法 2　求解类别近似质量约简.

输入　决策系统 DS,邻域半径参数 δ,类别标记 p.

输出　一个针对第 p 类标记的约简 red.

步骤 1　由(11)式计算 $\gamma(AT,X_p)$;

步骤 2　令 $red \leftarrow \varnothing$,$\gamma(red,X_p)=0$;

步骤 3　若 $\gamma(red,X_p)<(1-\varepsilon)\cdot\gamma(AT,X_p)$,则执行以下循环,否则转步骤 4;

(1) $\forall a \in AT-red$,计算属性 a 的重要度 $\mathrm{Sig}(a,red,X_p)$;

(2) 选择属性 b,满足 $\mathrm{Sig}(b,red,X_p)=\max\{\mathrm{Sig}(a,red,X_p),\forall a \in AT-red\}$,令 $red=red \cup \{b\}$;

(3) 由(11)式计算 $\gamma(red,X_p)$;

步骤 4　输出 red.

3　实验分析

为了验证类别近似质量约简的有效性,选取了 8 组 UCI 数据集,它们的基本信息如表 1 所列.实验中,使用欧氏距离构造样本之间的相似度矩阵,邻域半径参数 δ 分别设定为 0.05、0.1、0.15.在此基础上进行了 2 组实验,分别比较了两种算法在约简后所求得的近似质量以及约简的长度,算法 1 与算法 2 中所涉及的阈值 ε 设置为 0.05.2 组实验的实验环境皆为 PC 机,双核 2.30 GHz CPU,4 GB 内存,Windows7 操作系统,MATLAB R2014a 实验平台.

表 1　数据集描述

ID	数据集	样本数	属性	类别及类别中的样本数
1	Breast Cancer Wisconsin (Diagnostic)	569	30	$X_1(212)/X_2(357)$
2	Congressional Voting	435	16	$X_1(267)/X_2(168)$
3	Dermatology	366	34	$X_1(112)/X_2(61)/X_3(72)/X_4(49)/X_5(52)/X_6(20)$
4	Ionosphere	351	34	$X_1(225)/X_2(126)$
5	Molecular Biology(Promoter Gene Sequences)	106	57	$X_1(53)/X_2(53)$
6	Page Blocks Classification	5 473	9	$X_1(4913)/X_2(329)/X_3(28)/X_4(88)/X_5(115)$
7	Parkinson Multiple Sound Recording	1 208	26	$X_1(688)/X_2(520)$
8	Wine	178	13	$X_1(59)/X_2(71)/X_3(48)$

表 2 两种算法在近似质量上的对比

ID	类别	δ=0.05			δ=0.1			δ=0.15		
		原始数据	算法 1	算法 2	原始数据	算法 1	算法 2	原始数据	算法 1	算法 2
1	X_1	0.797 2	0.773 6	0.768 9	0.632 1	0.612 3	0.622 6	0.500 0	0.490 6	0.485 8
	X_2	0.764 7	0.723 9	0.739 5	0.299 7	0.280 1	0.285 7	0.047 6	0.036 4	0.047 6
2	X_1	0.895 1	0.876 4	0.861 4	0.764 0	0.752 8	0.752 8	0.662 9	0.662 9	0.662 9
	X_2	0.851 2	0.807 4	0.839 3	0.696 4	0.636 9	0.696 4	0.523 8	0.523 8	0.517 9
3	X_1	0.973 2	0.955 4	0.937 5	0.946 4	0.928 6	0.908 1	0.919 6	0.910 7	0.875 0
	X_2	0.737 7	0.704 9	0.721 3	0.541 0	0.508 2	0.541 0	0.262 3	0.245 9	0.262 3
	X_3	0.986 1	0.986 1	0.986 1	0.986 1	0.934 4	0.986 1	0.986 1	0.930 6	0.972 2
	X_4	0.734 7	0.653 1	0.734 7	0.428 6	0.428 6	0.408 2	0.163 3	0.142 9	0.163 3
	X_5	1.000 0	0.980 8	0.961 5	0.942 3	0.942 3	0.923 1	0.846 2	0.846 2	0.826 9
	X_6	0.950 0	0.850 0	0.950 0	0.900 0	0.800 0	0.900 0	0.900 0	0.750 0	0.900 0
4	X_1	0.915 6	0.915 6	0.915 6	0.737 8	0.737 8	0.706 7	0.546 7	0.537 8	0.528 9
	X_2	0.373 0	0.317 5	0.357 1	0.230 2	0.198 4	0.230 2	0.127 0	0.095 2	0.127 0
5	X_1	0.660 4	0.660 4	0.660 4	0.301 9	0.301 9	0.301 9	0.150 9	0.150 9	0.150 9
	X_2	0.377 4	0.377 4	0.377 4	0.188 7	0.188 7	0.188 7	0.018 9	0.018 9	0.018 9
6	X_1	0.152 2	0.145 3	0.149 2	0.016 9	0.016 9	0.016 9	0.000 0	0.000 0	0.000 0
	X_2	0.203 6	0.200 6	0.197 6	0.030 4	0.030 4	0.030 4	0.006 1	0.006 1	0.006 1
	X_3	0.392 9	0.321 4	0.392 9	0.285 7	0.285 7	0.285 7	0.178 6	0.178 6	0.178 6
	X_4	0.022 7	0.022 7	0.022 7	0.011 4	0.011 4	0.011 4	0.000 0	0.000 0	0.000 0
	X_5	0.078 3	0.052 2	0.078 3	0.026 1	0.026 1	0.026 1	0.026 1	0.026 1	0.026 1
7	X_1	0.143 9	0.139 5	0.139 5	0.024 7	0.024 7	0.024 7	0.008 7	0.008 7	0.008 7
	X_2	0.203 8	0.196 2	0.194 2	0.105 8	0.105 8	0.101 9	0.057 7	0.057 7	0.055 8
8	X_1	1.000 0	1.000 0	0.983 1	0.983 1	0.970 2	0.983 1	0.864 4	0.864 4	0.830 5
	X_2	0.845 1	0.788 7	0.845 1	0.732 4	0.690 1	0.732 4	0.647 9	0.647 9	0.647 9
	X_3	0.979 2	0.979 2	0.958 3	0.895 8	0.844 2	0.895 8	0.708 3	0.708 3	0.687 5

表 2 列出了利用 2 种不同的属性约简方法所得到的近似质量.观察表 2 可以发现,算法 1 无法保证每一个决策类别上的近似质量的误差在约简后依然保持在给定的阈值范围内.例如,Dermatology 数据集的类别 X_6,当 $\delta=0.05$、0.1、0.15 时,近似质量在约简的误差分别达到了 11%、11%以及 17%.而算法 2 却可以保证每个决策类别上的近似质量误差在约简后都保持在给定的阈值范围内.

进一步地,表 3 展示了在不同邻域半径参数下利用算法 1 和算法 2 所求得约简的长度.

观察表 3 可知,由算法 1 和算法 2 所得到的约简长度相差不大.结合表 2 可以得到如下结论.

(1) 对于某些决策类别,算法 1 可以使得其约简后的近似质量误差在阈值允许范围内,算法 2 却可以用更少的属性使其约简后的近似质量误差在阈值允许范围内.例如 Ionosphere 数据集,当 $\delta=0.05$、0.1、0.15 时,算法 1 可以使得类别 X_1 约简后的近似质量误差在阈值允许范围内,其约简长度分别为 32,33 和 33,而利用算法 2 所求得的约简长度分别为 26,30 和 31.

(2) 对于另一些决策类别而言,如果算法 1 不能使其约简后的近似质量误差在阈值允许范围内,那么算法 2 往往需要更多的属性来使得对应决策类别的近似质量误差在阈值允许范围内.例如 Ionosphere 数据集,当 $\delta=0.05$、0.1、0.15 时,算法 1 不能使得类别 X_2 约简后的近似质量误差在阈值允许范围内,其约简长度分别为 32,33 和 33,而利用算法 2 则可以使得类别 X_2 约简后的近似质量误差在阈值允许范围内,所对应的约简长度分别为 33,34 和 34.

表 3　约简长度的对比　　个

ID	算　法	类　别	$\delta=0.05$	$\delta=0.1$	$\delta=0.15$
1	算法 1		25	27	28
	算法 2	X_1	25	26	26
		X_2	26	28	30
2	算法 1		15	15	16
	算法 2	X_1	14	15	16
		X_2	15	16	15
3	算法 1		28	28	28
	算法 2	X_1	19	22	25
		X_2	21	26	26
		X_3	7	10	12
		X_4	19	25	30
		X_5	19	25	19
		X_6	32	25	27
4	算法 1		32	33	33
	算法 2	X_1	26	30	31
		X_2	33	34	34
5	算法 1		8	9	9
	算法 2	X_1	8	9	1
		X_2	6	8	7
		X_3	8	9	9
		X_4	7	3	1
		X_5	9	5	8
6	算法 1		57	57	57
	算法 2	X_1	57	57	57
		X_2	56	57	57
7	算法 1		25	26	26
	算法 2	X_1	25	26	25
		X_2	24	23	24
8	算法 1		12	12	13
	算法 2	X_1	10	12	11
		X_2	13	12	13
		X_3	11	13	12

4　结束语

在邻域粗糙集中,传统基于近似质量的属性约简仅考虑面向整体数据的近似质量变化,而忽略了具体某种决策类别所对应的近似质量变化.鉴于此,引入了基于类别近似质量的属性约简方法,并利用启发式算法求解约简.实验结果表明,该方法可以保证每一决策类别的近似质量都能够满足属性约简的约束条件.

在此基础上,下一步将讨论由不同决策类所生成类别近似质量约简之间的结构关系,同时亦可将类别近似质量约简引入到其他的粗糙集模型中去,以扩展类别近似质量约简的应用范围.

参 考 文 献

[1] Pawlak Z.Rough sets[J].International Journal of Computer and Information Sciences,1982,11(5):341-356.

[2] 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.

[3] Xu S P,Yang X B,Tsang Eric C C,et al.Neighborhood Collaborative Classifiers[C]// International Conference on Machine Learning and Cybernetics.South Korea:IEEE,2016:470-476.

[4] Yang X B,Chen Z H,Dou H L,et al.Neighborhood System Based Rough Set:Models and Attribute Reductions[J].International Journal of Uncertainty,Fuzziness and Knowledge-Based Systems,2012,20(3):399-419.

[5] Zhao H,Wang P,Hu Q H.Cost-sensitive Feature Selection Based on Adaptive Neighborhood Granularity with Multi-level Confidence[J].Information Sciences,2016,366:134-149.

[6] Wang C Z,Shao M W,He Q,et al.Feature Subset Selection Based on Fuzzy Neighborhood Rough Sets[J].Knowledge-Based Systems,2016,111:173-179.

[7] An S,Shi H,Hu Q H,et al.Fuzzy Rough Regression with Application to Wind Speed Prediction[J].Information Sciences,2014,282:388-400.

[8] 程晓荣,张兰,岳娇.基于粗糙集属性约简的评估模型在电力通信网风险评估中的应用及实现[J].电力系统保护与控制,2016,44(8):44-48.

[9] 王思华,杨桐,段启凡,等.基于 DT 法和粗糙集理论的接地网安全性状态评定[J].电力系统保护与控制,2017,45(2):48-54.

[10] Song J J,Tsang Eric C C,Chen D G,et al.Minimal Decision Cost Reduct in Fuzzy Decision-theoretic Rough Set Model[J].Knowledge-Based Systems,2017,124:104-112.

[11] 朱鹏飞,胡清华,于达仁.基于随机化属性选择和邻域覆盖约简的集成学习[J].电子学报,2012,40(2):273-279.

[12] Ju H R,Li H X,Yang X B,et al.Cost-sensitive Rough Set:A Multi-granulation Approach[J].Knowledge-Based Systems,2017,123:137-153.

[13] Yao Y Y,Zhang X Y.Class-specific Attribute Reducts in Rough Set Theory[J].Information Sciences,2017,418/419:601-618.

[14] 李智远,杨习贝,徐苏平,等.邻域决策一致性的属性约简方法研究[J].河南师范大学学报(自然科学版),2017,45(5):68-73.

[15] 杨习贝,颜旭,徐苏平,等.基于样本选择的启发式属性约简方法研究[J].计算机科学,2016,43(1):40-43.

[16] Xu S P,Yang X B,Yu H L,et al.Multi-label Learning with Label-specific Feature Reduction[J].Knowledge-Based Systems,2016,104:52-61.

[17] 魏巍,魏琪,王锋.粗糙集的不确定性度量比较研究[J].南京大学学报(自然科学版),2015,51(4):714-722.

[18] Mi J S,Wu W Z,Zhang W X.Approaches to Knowledge Reduction Based on Variable Precision Rough Set Model[J].Information Sciences,2004,159(3-4):255-272.

[19] Zhang X,Mei C L,Chen D G,et al.Feature Selection in Mixed Data:A Method Using a Novel Fuzzy Rough Set-based Information Entropy[J].Pattern Recognition,2016,56(1):1-15.

[20] 王宇,杨志荣,杨习贝.决策粗糙集属性约简:一种局部视角方法[J].南京理工大学学报(自然科学版),2016,40(4):444-449.

[21] Chen D G,Zhao S Y.Local Reduction of Decision System with Fuzzy Rough Sets[J].Fuzzy Sets and Systems,2010,161(13):1871-1883.

Attribute reduction constrained by class-specific approximate quality

Li Zhiyuan[1], Yang Xibei[2a], Chen Xiangjian[2a], Wang Pingxin[2b]

(1.Kewen College,Jiangsu Normal University,Xuzhou 221116,China;

2.a.School of Computer;b.School of Science,Jiangsu University of Science and Technology,Zhenjiang 212003,China)

Abstract: Based on the measurement of approximate quality, the traditional heuristic algorithm for computing reduction is designed to the find redundant attributes through considering the variation of approximate quality. However, such an approach does not take the variation of lower approximation of each decision class with reduction into account. To fill such a gap, a class-specific approximate-quality-based reduction is proposed. The objective of this strategy is to make the approximate quality of each decision class be acceptable in terms of the constraint of attribute reduction. By using the neighborhood rough set, traditional attribute reduction and class-specific approximate quality based strategies are compared over several UCI data sets. The experimental results tell us that not only the class-specific approximate quality based strategy is effective, but also it satisfies the constraint of traditional attribute reduction.

Keywords: attribute reduction; class-specific approximate quality; heuristic algorithm; rough set

[责任编校 陈留院]