文章编号：1004-4353(2018)02-0149-06

基于改进辨识矩阵的变精度邻域粗糙集属性约简

沈　林

（莆田学院 信息工程学院，福建 莆田 351100）

摘要：提出一种用于变精度邻域粗糙集，可以大幅减少时间复杂度的属性约简算法. 该算法基于一种改进的辨识矩阵. 首先用辨识矩阵同时记录决策一致和不一致的数据，然后用二进制位运算计算样本的邻域，最后获得可以保持下近似分布不变的属性约简. 实验结果证明，本文算法不仅能够大幅减少属性约简时间，而且精度上总体优于 NBRS 算法和 LDNRS 算法.

关键词：变精度邻域粗糙集；辨识矩阵；属性约简

中图分类号：TP181　　**文献标识码**：A

Attribute reduction of variable precision neighborhood rough sets based on improved identification matrix

SHEN　Lin

（*College of Information Engineering*, *Putian University*, *Putian* 351100, *China*）

Abstract: In this paper, an attribute reduction algorithm is proposed for variable precision neighborhood rough sets, which can greatly reduce the time complexity. The algorithm is based on an improved discernibility matrix. Firstly, consistent and inconsistent decision data is recorded by the matrix at the same time. Then, neighborhood of the sample is computed by binary bit operation. Finally, an attribute reduction that can keep the lower approximation distribution unchanged can be obtained. Experimental results show that this algorithm can greatly reduce the needed time of attribute reduction, and is generally better than NBRS and LDNRS in accuracy.

Keywords: variable precision neighborhood rough sets; identification matrix; attribute reduction

1991 年，Z.Pawlak[1] 提出了粗糙集理论(简称 RS). RS 基于等价关系构建了上近似集和下近似集，分别表示模糊知识和精确知识. 通过构建上下近似集，RS 可以从不确定、不完备以及不一致决策表中挖掘出潜藏的知识，但是 RS 不适合处理连续型数据，并且对抗噪声能力弱. 为此，W.Ziarko[2] 提出了变精度粗糙集(VPRS)，将精度超过一个阀值的等价关系归为一类，改善了抗噪音能力；HU 等[3-4] 提出了邻域粗糙集(NBRS)，用邻域关系代替等价关系处理连续型数据，并对变精度邻域粗糙集进行了研究，但作者并未考虑每个决策类是否都可以达到较高的精度；沈林等[5] 在 HU 等研究的基础上将下近似分布不变引入到变精度邻域粗糙集(LDNRS)，该方法能够保证整体精度较高，避免出现个别决策类精度较低的情况.

数据集的属性约简(以下简称为约简)是粗糙集最重要的应用之一，主要分为基于辨识矩阵和基于

收稿日期：2018－03－08　　**基金项目**：福建省教育厅项目(JA15458)

作者简介：沈林(1983—)，男，讲师，研究方向为人工智能、机器学习、粗糙集理论.

依赖度两种方法.其中辨识矩阵无须求出所有的最小约简,因此具有更快的约简速度.VPRS 在使用辨识矩阵时主要有两种方法:一是将决策不一致率小于阀值的等价关系视为决策一致,该方法改变了数据的含义[6];另一种是将决策一致和不一致的等价关系分开处理[7].由于 NBRS 是基于邻域关系,上述两种方法均无法用于 NBRS,因此目前基于辨识矩阵的 NBRS 约简研究主要针对决策一致数据.文献[8]和文献[9]分别将传统辨识矩阵和二进制辨识矩阵应用于 NBRS,但这两种方法均无法处理决策不一致数据.基于此,本文提出一种改进的辨识矩阵.该方法同时记录决策一致和不一致的信息,并利用二进制位运算快速计算样本的邻域,减少时间复杂度,同时要求约简后保持下近似分布不变.利用 UCI 数据集对本文方法进行验证表明,本文方法不仅精度上总体优于 NBRS 算法和 LDNRS 算法,还能够大幅减少运行时间.

1 基本概念

1.1 变精度邻域粗糙集模型

设 $DS=(U,C\cup D,V,f)$ 为一决策系统,$U=\{x_1,x_2,\cdots,x_n\}$ 为非空样本集,C 和 D 分别是条件属性和决策属性,且 $C\cap D=\varnothing$,V 为 $C\cup D$ 的值域,f 是 $U\times(C\cup D)\to V$ 的映射.

定义 1 样本 x_i 的邻域关系记为 $\delta_A(x_i)=\{x_j\mid x_j\in U,\Delta_A(x_i,x_j)\leqslant\delta\}$,其中 δ 为邻域半径,$A\subseteq C$,$\Delta_A(x_i,x_j)$ 表示在属性 A 下样本 x_i 和 x_j 的距离.

根据定义 1 可知,必然有 $x_i\in\delta(x_i)$.若 $x_i\in\delta(x_j)$,则必然 $x_j\in\delta(x_i)$,但是无法从 $x_i\in\delta(x_j)$ 和 $x_k\in\delta(x_j)$ 中推导出 $x_k\in\delta(x_i)$.所以邻域关系满足自反性、对称性和非传递性.

定义 2 对于给定的信息系统 $DS=(U,C\cup D,V,f)$,引入错误率 $\beta(0\leqslant\beta\leqslant0.5)$,有非空样本集 $X\subseteq U$,则 X 关于属性 C 的 β 上、下近似可以描述为:

$$\bar{R}_{\beta\delta}(X)=\bigcup\{x_i\mid 1-\frac{\delta_C(x_i)\cap X}{\delta_C(x_i)}\leqslant 1-\beta,x_i\in U\},$$

$$\underline{R}_{\beta\delta}(X)=\bigcup\{x_i\mid 1-\frac{\delta_C(x_i)\cap X}{\delta_C(x_i)}\leqslant \beta,x_i\in U\}. \tag{1}$$

定义 3 在决策系统 DS 中,有属性 $A\subseteq C$.若有 $x_i,x_j\in U$ 在属性 A 下互为邻域,则当 x_i 和 x_j 的决策属性一致时,称 x_i 和 x_j 决策一致,否则称为决策不一致.

1.2 改进的辨识矩阵

本文用一个辨识矩阵来描述定义 1 中的邻域关系,该矩阵每行为由两个样本构成的样本对,每列表示一个属性,具体定义如定义 4.

定义 4 对于任意的 $x_i,x_j\in U$ 及属性 $a\in C$,决策系统 DS 的辨识矩阵 $M(ij)a$ 定义为:

$$M(ij)a=\begin{cases}2, & x_j\in\delta_a(x_i)\wedge f(x_i,D)\neq f(x_j,D)\wedge i<j;\\ 1, & x_j\in\delta_a(x_i)\wedge f(x_i,D)=f(x_j,D)\wedge i<j;\\ 0, & \text{其他}.\end{cases} \tag{2}$$

根据定义 4 知,在属性 a 下,若样本 x_i 和 x_j 是非邻域关系,则 $M(ij)a$ 值为 0;若 x_i 和 x_j 是邻域关系且决策一致,则 $M(ij)a$ 值为 1;若 x_i 和 x_j 是邻域关系但决策不一致,则 $M(ij)a$ 值为 2.根据定义 3 可知,矩阵中任意一行均不会同时出现 1 和 2.若要计算样本 x_i 和 x_j 在属性 $a_1\cup a_2$ 下是否互为邻域关系,只需进行 $M(ij)a_1\&M(ij)a_2$ 运算.$1\&1\Rightarrow1$,说明在属性 $a_1\cup a_2$ 下是邻域关系且决策一致;$2\&2\Rightarrow2$,说明是邻域关系但决策不一致;$0\&(0\,||\,1\,||\,2)\Rightarrow0$,说明不是邻域关系.整个矩阵列数等于数据集条件属性 C 的个数,行数为 $m(m-1)/2$,m 为样本个数.

1.3 下近似分布相关定义

定义 5 决策系统基于 β 的近似邻域依赖:

$$r(DS,\beta)=|POS(DS,\beta)|/|U|. \tag{3}$$

其中 $POS(DS,\beta)=\bigcup C_{\beta\delta}Y_j$，$Y_j\subseteq U/D$ 是决策属性 D 对样本 U 的划分，$C_{\beta\delta}Y_j$ 为决策类 Y_j 在条件属性 C 的 δ 邻域关系下的 β 下近似，$POS(DS,\beta)$ 是决策类下近似的并集.

对于属性集 $A\subset C$，当 $r_A(DS,\beta)=r(DS,\beta)$ 时，则认为属性 A 是 C 的一个约简. 但在引入错误率 β 后，有可能会出现系统整体的依赖度不变，但某些决策类的样本会发生变化，这会增大决策错误的可能性. 为此，可以要求在约简时保持下近似分布不变，以解决该问题[5].

定义 6　决策系统的下近似分布的定义为：

$$DP(DS,\beta)=\{C_{\beta\delta}Y_1,C_{\beta\delta}Y_2,\cdots,C_{\beta\delta}Y_n\}. \tag{4}$$

定义 7　若有属性集 $A\subset C$，且属性集 A 的下近似分布 $DP_A(DS,\beta)=DP(DS,\beta)$，则认为属性集 A 是 C 的一个下近似分布约简.

1.4　辨识矩阵的建立

下面通过具体的实例来说明如何建立本文所提出的辨识矩阵.

例 1　表 1 为一个包含 4 个条件属性及 1 个决策属性的数据集，样本为 5 个，所有数据均为连续型数据.

表 1　决策信息表

样本	a_1	a_2	a_3	a_4	d
x_1	0.52	0.12	0.55	0.36	1
x_2	0.64	0.27	0.62	0.50	0
x_3	0.60	0.39	0.60	0.45	0
x_4	0.21	0.36	0.23	0.10	1
x_5	0.63	0.16	0.63	0.49	1

由于每个属性的数据分布特性不同，所以应分别设置邻域半径. 本文取标准差的 1/2 作为邻域半径，各属性的邻域半径如表 2 所示. 根据表 1、表 2 和定义 4，计算得辨识矩阵，如表 3 所示.

表 2　表 1 中条件属性的邻域半径

	a_1	a_2	a_3	a_4
邻域半径	0.09	0.06	0.08	0.08

表 3　表 1 的邻域辨识矩阵

样本对	辨识矩阵				样本对	辨识矩阵			
	a_1	a_2	a_3	a_4		a_1	a_2	a_3	a_4
(x_1,x_2)	0	0	2	0	(x_2,x_4)	0	0	0	0
(x_1,x_3)	2	0	2	0	(x_2,x_5)	2	0	2	2
(x_1,x_4)	0	0	0	0	(x_3,x_4)	0	2	0	0
(x_1,x_5)	0	1	1	0	(x_3,x_5)	2	0	2	2
(x_2,x_3)	1	0	0	1	(x_4,x_5)	0	0	0	0

2　基于位运算辨识矩阵的变精度邻域粗糙集属性约简算法

邻域辨识矩阵的每一行均存储了两个样本在各属性下的邻域关系及决策是否一致的信息，根据定义 4 对辨识矩阵进行位运算，就可以实现各属性组合下样本的邻域计算、错误率计算. 文献[5]在测试候选属性组合时，需要重新计算每个样本的邻域错误率，导致该算法耗时过多. 本文在每一轮属性选择中，选择可以使决策不一致样本对比例最低的属性，即数字 2 比例最低的属性，无需反复统计各个样本的错误率，因此降低了时间复杂度. 本文算法同时要求获得的约简可以保持下近似分布不变，以保证每个决策类都可以获得符合要求的精度. 将本文算法简称为 BMLNRS，其约简流程如下：

输入：决策系统 $DS=(U,C\cup D,V,f)$，邻域半径 δ，错误率 β.

输出：约简后的属性集.

1) 按照定义 4,计算数据集的邻域辨识矩阵；

2) 计算全属性 C 下的下近似分布；

3) 找出最小错误率的属性$\{a_i \mid \min(\mid M(ij)a_i=2\mid /(\mid M(ij)a_i=1\mid+m))\}$，$m$ 是元素个数，并将该属性放入已选属性队列，然后执行步骤 5)；

4) 将剩余属性依次和已选属性队列做二进制与运算，将最小错误率的属性加入已选属性队列，若有多个剩余属性可以得到最小错误率，选择数值 1 最多的剩余属性；

5) 检查下近似分布是否和 2) 一致，如果是则输出已选属性队列并结束算法，如果不是则重复 4)、5) 步骤，直到满足条件.

整个算法中步骤 4) 耗时最多，时间复杂度为 $O(m^2*n*l)$，m 为数据集中的样本个数，n 为条件属性个数，l 为约简后的属性个数.

以表 3 的辨识矩阵为例，对表 3 的所有列进行二进制与运算后，可以求得各个样本的邻域关系：$\delta_c(x_1)=\{x_1\}$、$\delta_c(x_2)=\{x_2\}$、$\delta_c(x_3)=\{x_3\}$、$\delta_c(x_4)=\{x_4\}$、$\delta_c(x_5)=\{x_5\}$. 依据定义 6，可以求得全属性 C 时下近似分布为$\{(x_1,x_4,x_5),(x_2,x_3)\}$. 接下来按照步骤 3) 中的公式计算，发现属性 a_2 的错误率最低，为 0.17，所以先将 a_2 放入已选属性队列. 此时下近似分布为$\{(x_1,x_5),(x_2)\}$，和全属性 C 时的不一致. 继续计算发现，组合(a_2,a_1)、(a_2,a_3)、(a_2,a_4) 的错误率均为 0，其中组合(a_2,a_3) 的 1 的个数最多，所以应当选择(a_2,a_3) 作为约简.

3 实验分析

选取文献[3]的 NBRS 算法、文献[5]的 LDNRS 算法同本文的 BMLNRS 算法进行效率及精度对比，3 种算法均采用 Matlab 编写，运行环境为 I5-3470，16 G 内存. 为了保证算法的有效性和公正性，本文从 UCI 数据集中选择多个常用的数据集用于验证算法，并通过 WEKA 自带的 KNN、NaiveBayes 以及 SimpleCart 算法进行精度验证.

分别测试了 1/4、1/2 和 3/4 标准差下，错误率为 0.1、0.2、0.3、0.4 和 0.5 时的情况. 在测试中发现，约简后属性的个数随着邻域半径的增加而增加，这是因为邻域半径增加后，样本邻域中决策不一致的样本比例增加所致，因此必须引入更多的属性来稀疏样本的分布. 同时还发现，错误率的改变对约简后属性的个数影响较小. 图 1 和图 2 分别为 wine、ionosphere 数据集在不同邻域半径和错误率下的约简情况. 图 3 和图 4 分别为 wine、ionosphere 数据集在不同邻域半径下约简后精度变化的情况.

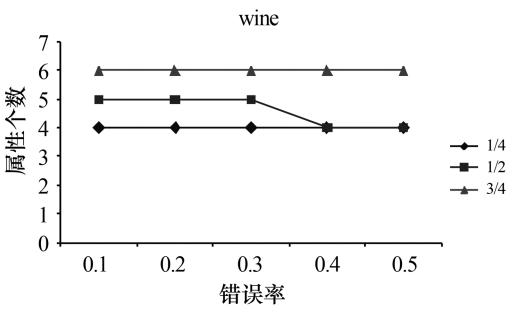

图 1 Wine 数据集的约简效果

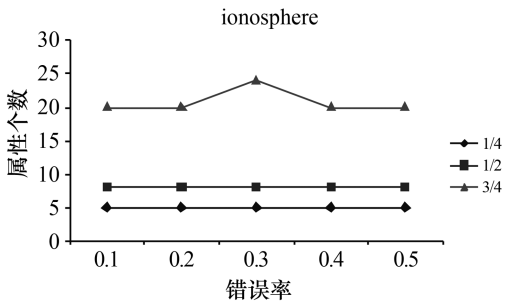

图 2 Ionosphere 数据集的约简效果

由图 2 可知，当邻域半径选择 3/4 标准差时，若数据集属性较多，容易发生约简后属性个数较 1/2 标准差激增的情况. 邻域半径选择 1/4 标准差时，在表 4 所列数据集的测试结果中，绝大多数情况下得

到的精度低于 1/2 标准差，因此邻域半径选择 1/2 标准差比较合适. 表 4—表 7 列举的数据均是在错误率为 0.3、邻域半径为 1/2 标准差时采集的.

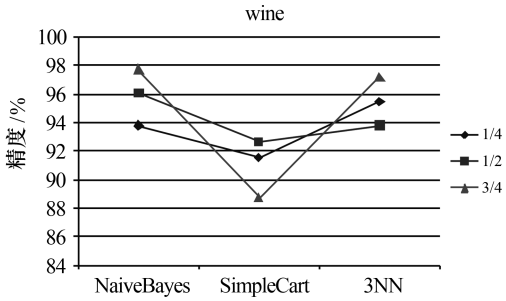

图 3　Wine 数据集的约简效果

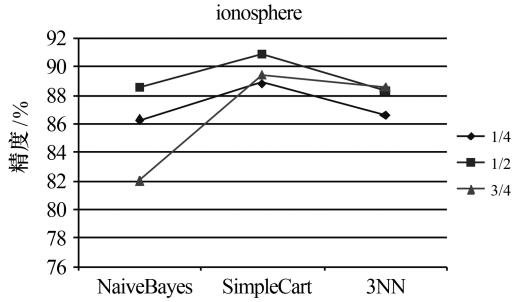

图 4　Ionosphere 数据集的约简效果

表 4　3 种算法的约简效果及时间对比

数据集	属性个数	样本个数	BMLNRS		LDNRS		NBRS	
			约简后属性个数	时间/s	约简后属性个数	时间/s	约简后属性个数	时间/s
wine	13	178	5	0.1	6	1.2	5	1.2
ionosphere	34	351	8	0.8	10	50	4	14.7
glass	9	214	8	0.2	7	2.6	7	1.5
wdbc	31	569	8	2.4	8	102	7	19.6
zoo	16	101	5	0.05	5	0.9	5	0.2
mushroom	22	8 124	4	178	3	>15 h	3	1 771
diabetes	8	768	8	1.4	8	31.3	8	10.2

从表 4 可以看出，3 种方法都可以对除 glass 和 diabetes 外的大部分数据集进行有效地约简，其中 BMLNRS 和 LDNRS 算法约简后的属性个数基本相同，说明约简能力大致相同. NBRS 约简算法则在 ionosphere 数据集下仅获得了 4 个属性的约简，大幅少于 BMLNRS 和 LDNRS 算法. 在运算时间上，BMLNRS 算法的运行时间最少，显著低于 LDNRS 和 NBRS 算法的运行时间. 这是因为本文的 BMLNRS 算法因采用了基于二进制位运算的改进邻域辨识矩阵，无需直接计算邻域和邻域错误率，所以大幅减少了运算时间.

在 diabetes 数据集下，3 种方法约简后均没有减少属性的个数，所以本文在以下内容中不再测试该数据集. 由表 5—表 7 可以看出：用 3NN 检验精度时，本文算法(BMLNRS)所得的精度略低于 NBRS 算法和 LDNRS 算法，但相差不大；用 NaiveBayes 和 SimpleCart 检验精度时，本文算法平均精度最好. 由此可以看出，本文算法的精度总体优于 NBRS 算法和 LDNRS 算法.

表 5　NBRS 算法的分类精度

数据集	属性个数	分类精度/%		
		NaiveBayes	SimpleCart	3NN
wine	5	93.82±4.6	88.76±9.1	96.63±3.3
ionosphere	4	81.77±24.4	88.89±16.9	86.32±18.4
glass	7	46.73±19	69.63±13.6	69.6±11.7
wdbc	7	94.2±6.5	93.85±8	95.78±6.21
zoo	5	95.05±2.7	95.05±1.9	90.1±3
mushroom	3	98.74±2.1	99.41±1.2	99.27±0.6
Average	5.17	85.05±9.9	89.27±8.5	89.62±7.2

表 6 LDNRS 算法的分类精度

数据集	属性个数	分类精度/%		
		NaiveBayes	SimpleCart	3NN
wine	6	95.51±3.9	90.45±7.5	94.94±4.4
ionosphere	10	88.6±12.2	92.3±10.8	90.6±11.2
glass	7	48.13±18.6	64.49±14.2	68.22±11.7
wdbc	8	94.2±7.3	93.15±8.6	95.08±6.3
zoo	5	97.03±2.1	94.06±3	88.12±3.2
mushroom	3	99.7±2.1	99.7±0.5	99.7±0.5
Average	6.5	85.53±7.7	89.03±7.4	89.44±6.22

表 7 BMLNRS 算法的分类精度

数据集	属性个数	分类精度/%		
		NaiveBayes	SimpleCart	3NN
wine	5	96.07±3.7	92.67±6.8	93.82±5.6
ionosphere	8	88.6±12.7	90.89±11.9	88.32±13.2
glass	8	46.26±18.4	64.49±14.6	69.63±11.6
wdbc	8	95.96±5.2	95.25±7.1	95.43±6
zoo	5	97.03±2.1	94.06±3	88.12±3.2
mushroom	4	98.92±2.2	99.9±0.2	99.9±0.2
Average	6.3	87.14±7.4	89.54±7.3	89.2±6.6

4 结束语

本文提出了一种改进的基于二进制位运算的辨识矩阵,经采用 UCI 数据集实验证明,本文算法在时间复杂度上显著优于文献[3]中的 NBRS 方法和文献[5]中的 LDNRS 方法,且在用 NaiveBayes 和 SimpleCart 检验精度时,其精度总体上优于 NBRS 算法和 LDNRS 算法,因此本文方法具有较好的应用价值,适用于对属性较多或样本较多的数据集进行属性约简.因本文算法的辨识矩阵中的每一行均为一个样本对的辨识信息,因此矩阵行数较多,其中包含无须计算的冗余样本对,因此在今后的工作中将探讨如何通过压缩矩阵规模,以进一步压缩算法的运行时间.

参考文献:

[1] Pawlak Z. Rough-Sets: Theoretical Aspects of Reasoning About Data[M]. Dordrecht: Kluwer Academic Publisher, 1991.

[2] Ziarko W. Variable precision rough set model[J]. Journal of Computer System Science, 1993,46(1):39-59.

[3] Hu Qinghua, Yu Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008,19(3):640-649.

[4] Hu Qinghua, Zhao Hui, Yu Daren. Efficient symbolic and numerical attribute reduction with neighborhood rough sets[J]. PR & AI, 2008,21(6):732-738.

[5] 沈林,陈建辉.基于下近似分布的变精度邻域粗糙集属性约简算法[J].贵州大学学报(自然科学版),2017,34(4):53-58.

[6] 杨燕燕.变精度粗糙集属性约简理论与算法[D].北京:华北电力大学,2013.

[7] 李艳,郭娜娜,赵浩.基于变精度和浓缩布尔矩阵的属性约简[J].计算机科学,2017,44(6A):70-74.

[8] 林俊伟,叶东毅.基于邻域辨识矩阵的属性约简增量式算法[J].计算机应用,2009,29(6):119-121.

[9] 杨云霞,杨占勇.二进制分辨矩阵在连续属性约简中的研究[J].计算机与数字工程,2012,40(1):19-24.