文章编号：1004-4353(2019)01-0049-06

邻域粗糙集的随机集成属性约简

沈　林

（莆田学院 信息工程学院，福建 莆田 351100）

摘要：为了解决传统的辨识矩阵空间复杂度高，难以应用于大规模数据的问题，提出了一种基于随机抽样的属性约简算法. 首先随机抽取多个小样本子集，以降低辨识矩阵的空间复杂度；然后分别对每个样本子集进行属性约简，并计算每个属性子集的权重；最后选择高权重的几个属性子集进行测试，找出精度最高的属性子集. 实验结果证明，本文方法比传统辨识矩阵的占用空间降低 2～3 个数量级，并且精度与其基本相当.

关键词：邻域粗糙集；随机抽样；属性约简；集成

中图分类号：TP181　　**文献标识码**：A

Random ensemble attribute reduction of neighborhood rough sets

SHEN　Lin

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

Abstract: In order to solve the problem that the traditional identification matrix has high spatial complexity, and is difficult to be applied to large-scale data, an attribute reduction algorithm based on random sampling is proposed. Firstly, several small sample subsets are randomly extracted to decrease the spatial complexity of the identification matrix; Secondly, attribute reduction is performed for each sample subset, and the weight of each attribute subset is calculated. Finally, several attribute subsets with high weights are selected for testing to find out the most accurate attribute subset. The experimental results show that the proposed method can reduce the occupied space by 2 to 3 orders of magnitude than traditional identification matrix, and its accuracy is basically the same as that of the traditional identification matrix.

Keywords: neighborhood rough sets; random sampling; attribute reduction; ensemble

2008 年，胡清华等通过引入邻域关系，构建了邻域粗糙集模型(neighborhood rough sets，NRS)[1]，解决了 Pawlak 的粗糙集理论(rough sets，RS)无法处理数值型数据的问题[2]. 此后，许多学者对邻域粗糙集模型进行了研究[3-6]. 目前，NRS 被广泛应用于知识发现、规则提取等领域. 属性约简是去除冗余属性、获得精简知识的前提，对研究 NRS 具有重要作用. 辨识矩阵是粗糙集属性约简的主要方法之一，但由于传统辨识矩阵仅记录样本间的辨识信息，缺少决策分布的相关信息，因此难以应用于变精度邻域粗糙集(variable precision neighborhood rough sets，VPNRS)的属性约简. 文献[7]提出了一种改进的辨识矩阵，解决了变精度邻域粗糙集的属性约简问题，但该改进的辨识矩阵占用空间较大，限制了其在大规模数据上的应用. 基于上述研究，本文提出一种基于随机抽样的属性约简算法，并通过对多个 UCI 数据集的实验，验证本文方法的可行性.

收稿日期：2019－01－21　　**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.

1 基本概念

1.1 邻域粗糙集和变精度邻域粗糙集模型

定义 1[1] 有决策系统 $DS=(U,C\cup D,V,f)$,U 是非空样本集,C 是条件属性,D 是决策属性,f 是 $U\times(C\cup D)\to V$ 的映射函数.样本 $x_i\in U$ 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 是邻域半径,属性集 $A\subseteq C$,$\Delta_A(x_i,x_j)$ 是样本 x_i 和 x_j 的距离函数.

定义 2[1] 对于给定的集合 $X\subseteq U$,属性集 A 的上近似和下近似定义为:

$$\begin{aligned}&\overline{R}_\delta(X)=\{x_i\mid\delta_A(x_i)\cap X\neq\varnothing,x_i\in U\};\\&\underline{R}_\delta(X)=\{x_i\mid\delta_A(x_i)\subseteq X,x_i\in U\}.\end{aligned}\tag{1}$$

上下近似是粗糙集中的重要的概念之一,是用于分析精确、模糊知识的重要工具.定义 2 要求处理的样本必须是精确的,但因其抗噪音能力差,所以在实践中往往会引入精度 $\beta(0\leqslant\beta\leqslant0.5)$,即将粗糙集变为变精度邻域粗糙集.变精度邻域粗糙集的上下近似定义为:

定义 3[1] 当有非空样本集 $X\subseteq U$,则 X 关于属性 A 的 β 上、下近似可以描述为:

$$\begin{aligned}&\overline{R}_{\beta\delta}(X)=\bigcup\{x_i\mid\frac{\delta_A(x_i)\cap X}{\delta_A(x_i)}\geqslant\beta,x_i\in U\};\\&\underline{R}_{\beta\delta}(X)=\bigcup\{x_i\mid\frac{\delta_A(x_i)\cap X}{\delta_A(x_i)}\geqslant1-\beta,x_i\in U\}.\end{aligned}\tag{2}$$

1.2 基于依赖度的属性约简

基于依赖度属性约简的基本思想是通过计算依赖度,寻找到可以保持正域不变的属性约简(下面记作 Dependence 算法).

定义 4[2] 决策系统的近似邻域依赖为

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

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

定义 5[2] 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.

1.3 基于辨识矩阵的变精度邻域粗糙集属性约简

因传统的辨识矩阵不能直接应用于 VPNRS,文献[8]定义了一种新的辨识矩阵,如公式(4)所示:

$$M(ij)a=\begin{cases}2,&x_j\in\delta_a(x_i)\wedge f(x_i,D)\neq f(x_j,D)\wedge(i<j);\\1,&x_j\in\delta_a(x_i)\wedge f(x_i,D)=f(x_j,D)\wedge(i<j);\\0,&\text{其他}.\end{cases}\tag{4}$$

该辨识矩阵的每行是一个样本对,每列对应一个属性,数字 0 表示样本对不是邻域关系,数字 1 表示样本对是邻域关系且决策属性相同,数字 2 表示样本对是邻域关系但决策属性不同.在每一轮属性选择中,选择数字值 2 与数字值 1 比值最低的属性.由于该方法无需反复计算各样本的邻域和精度,因此降低了时间复杂度.为了避免对某个决策类过度拟合,文献[7]的算法在约简过程中还检验了下近似分布不变.

定义 6[7] 决策系统的下近似分布的定义为:

$$DP(DS,\beta)=\{C_{\beta\delta}Y_1,C_{\beta\delta}Y_2,\cdots,C_{\beta\delta}Y_n\},\tag{5}$$

其中 $C_{\beta\delta}Y_j$ 为决策类 Y_j 在条件属性 C 的 δ 邻域关系下的 β 下近似.

文献[7]中的改进辨识矩阵算法(下面称为 BMLNRS 算法)的具体流程如下:

a) 按照式(3)计算样本集的邻域辨识矩阵.

b) 计算全属性 C 下的下近似分布.

c) 找出精度最高的属性$\{a_i \mid \min(\mid M(ij)a_i=2\mid /(\mid M(ij)a_i=1\mid +m))\}$，$m$ 是元素个数，并将该属性放入已选属性队列，然后执行步骤 e).

d) 将剩余属性依次和已选属性队列做位与运算，将精度最高的属性加入已选属性队列. 若有多个剩余属性可以得到最高精度，则选择数值 1 最多的剩余属性.

e) 检查下近似分布是否和 b) 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 d)、e) 步骤，直到满足条件.

2　基于随机抽样的集成属性约简

为了解决 BMLNRS 算法空间占用过高的问题，本文通过随机抽样获得多个不同的小规模样本，然后利用 BMLNRS 算法分别进行约简. 在获得多个有一定差异的属性子集后，计算每个属性子集的权重，并选取最好的 n 个属性子集在之前抽取的小规模样本上进行测试，以此选出精度最好的属性子集. 为了进一步减少空间占用，将文献[7] 中按字节存储矩阵元素的方法，改为按二进制位存储. 整个算法如图 1 所示.

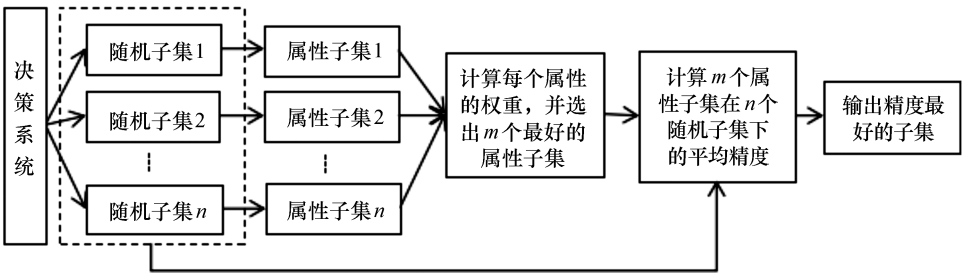

图 1　本文算法的流程图

在计算属性子集的权重时，若在多组不同的属性子集中，某属性出现的次数多，则表示其分辨决策类的能力强. 权重的计算公式如下：

$$\omega_{C_i}=\sum_{i=1}^{n}\frac{1}{\mathrm{MAX}(\mid S_i\mid /\mid S_{\mathrm{avg}}\mid ,\mid S_{\mathrm{avg}}\mid /\mid S_i\mid)\cdot \mid S_i\mid },$$

$$\omega_{S_i}=\Big(\sum_{j=1}^{n}\omega_{C_j}\mid C_j\in S_i\Big)/\mathrm{MAX}(\mid S_i\mid /\mid S_{\mathrm{avg}}\mid ,\mid S_{\mathrm{avg}}\mid /\mid S_i\mid), \tag{6}$$

其中 ω_{C_i} 表示属性 C_i 的权重，ω_{S_i} 表示属性子集 S_i 的权重.

3　实验结果

所有实验均在 Windows7 环境下完成. 首先在 Matlab 下编写代码，以此获得属性约简的子集，然后使用 WEKA 自带的算法验证精度. 本文从 UCI 数据集中选择 5 个大规模数据集来验证本文算法的效果，所有数据集均为数值型数据，如表 1 所示.

表 1　数据集参数

数据集名称	缩写	样本个数	属性个数	决策类别个数
MAGIC Gamma Telescope	Magic	19 020	10	2
EEG Eye State	EEG	14 980	14	2
Statlog (Shuttle)	Shuttle_tst	14 500	9	5
Letter Recognition	Letter	20 000	16	26
Waveform Database Generator (Version 2)	Waveform	5 000	40	3

3.1 各抽样比例的空间占用和时间消耗

不同抽样比例对辨识矩阵空间占用的影响如表 2 所示. 从表 2 可以看出,随着抽样比例的降低,辨识矩阵的占用空间迅速减少. 在 30%抽样比例时,占用空间为全集的 2%～3%;在 10%抽样比例时,占用空间只有全集的 0.25%～0.35%. 这说明,随机小规模样本子集可以显著减少辨识矩阵的占用空间.

表 2 各抽样比例的空间占用

数据集	占用空间/MB			
	全集	30%	20%	10%
Magic	1 725	46.5	20.7	5.2
EEG	1 498	38.5	17.1	4.3
Shuttle_tst	902	27	12	3
Letter	3 051	68.6	30.5	7.6
Waveform	476	10.7	4.7	1.2

为了避免算法运行时间超过全集时的运行时间,将本文算法的运行时间与全集时的 Dependence 算法、BMLNRS 算法进行比较,结果如表 3 所示. 随机抽取本文算法的 15 个样本子集(按 30%、20%、10%比例分别随机抽取 5 次)进行运行. 从表 3 可以看出,采用 15 组随机子集进行属性约简,其运行总时间明显少于 BMLNRS 算法和基于依赖度的算法. 这说明,通过控制随机抽样的次数,可以使属性约简的时间消耗不超过全集下属性约简的时间.

表 3 3 种算法的时间消耗

数据集	时间消耗/s		
	Dependence 算法	BMLNRS 算法	本文算法
Magic	26 697	4 767	3 840
EEG	14 108	4 793	4 555
Shuttle_tst	3 343	3 033	2 390
Letter	11 329	9 732	6 915
Waveform	3 150	1 451	900

3.2 各抽样比例和邻域半径下的稳定性

3.2.1 各抽样比例对约简子集的属性个数的影响 表 4 和表 5 给出了不同抽样比例对约简后属性个数的影响. 由表 4 可知,在 0.5σ 时,除 Waveform 外,其他数据集在各抽样比例下其约简后的属性个数与全集基本相当,即并不随抽样比例的变化而发生显著变化. 但在 0.3σ 时(表 5),各数据集约简后的属性个数均随抽样比例的降低而减少.

表 4 0.5σ 邻域半径下约简后的属性个数

数据集	全集约简属性个数	不同抽样比例约简属性个数			
		30%	20%	10%	AVERAGE
Magic	8	7.8	7.8	7.8	7.8
EEG	9	9.2	9.0	9.4	9.2
Shuttle_tst	6	6.2	6.0	6.0	6.1
Letter	10	9.0	9.2	8.8	9.0
Waveform	8	6.0	6.2	5.8	6.0

表 5 0.3σ 邻域半径下约简后的属性个数

数据集	全集约简属性个数	不同抽样比例约简属性个数			
		30%	20%	10%	AVERAGE
Magic	8	7.4	7.2	7.0	7.2
EEG	7	7.0	7.0	6.4	6.8
Shuttle_tst	5	4.8	4.2	4.4	4.5
Letter	6	5.4	5.2	5.2	5.3
Waveform	6	5.0	4.4	4.0	4.5

3.2.2　各抽样比例约简结果的相似度　为了进一步了解随机抽样和邻域半径对约简后属性子集的影响，将本文算法的 15 个随机样本子集分别按照 0.3σ、0.4σ、0.5σ、0.6σ 的邻域半径进行属性约简，并分别计算这些结果与全集约简结果的相似度，然后将相似度按照样本抽样比例分组并求平均值，结果如图 2 所示. 相似度计算采用谷元距离度量法[8]：

$$D_T = 1 - \frac{|S| + |S'| - 2|S \cap S'|}{|S| + |S'| - |S \cap S'|}. \tag{7}$$

公式中，D_T 的取值范围为[0,1]，取值越大，说明两个属性子集的相似度越高；取值为 0 时表示完全不同，为 1 时表示完全相同.

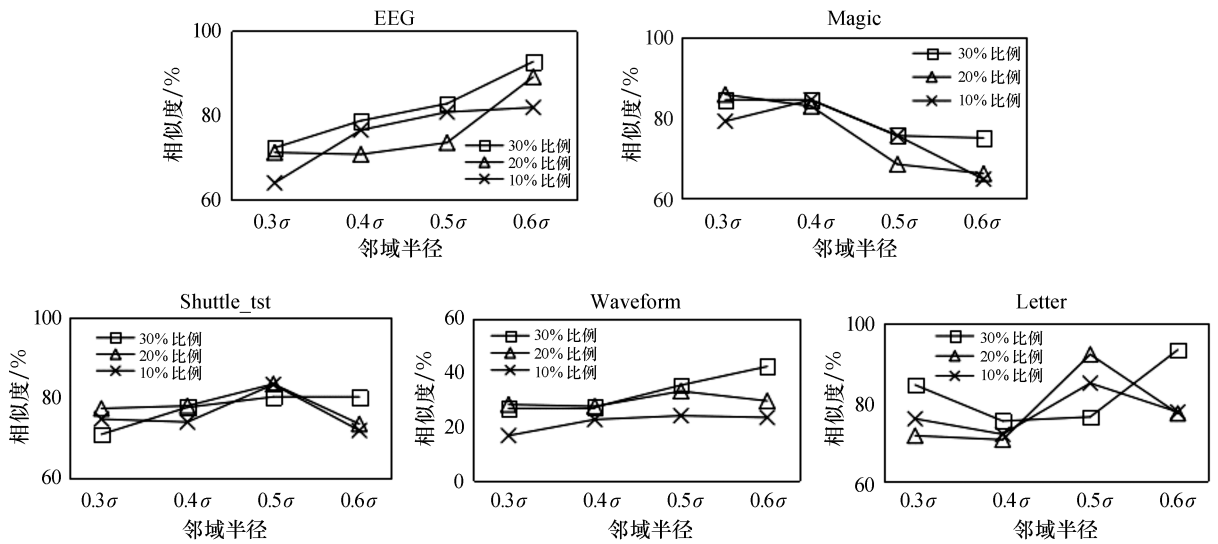

图 2　各抽样比例约简结果和全集约简结果的平均相似度

从图 2 可以看出，相似度的变化与抽样比例、邻域半径的变化没有相关性. 其中，Waveform 数据集的相似度低于其他数据集，这是因为 Waveform 数据集中有 40 个属性，而每个抽样比例仅随机抽取 5 个样本，所以相似度偏低.

3.2.3　各抽样比例的约简精度　按 30%、20%、10%比例随机抽样(每个比例各抽样 5 次)测试各抽样比例对精度的影响. 测试时，δ 取 0.5σ，β 取 0.5. 约简后，用 3NN、SimpleCart、SMO、Bagging、JRip、RandomForest 算法计算每组的平均精度，结果如图 3 所示. 由图 3 可以看出，在 30%和 20%的抽样比例下，除了 Letter 数据集，其他随机子集的精度都略高于全集. 在抽样比例为 10%时，随机子集的精度普遍较低，其原因是在该抽样比例下，样本子集的信息量丢失较多.

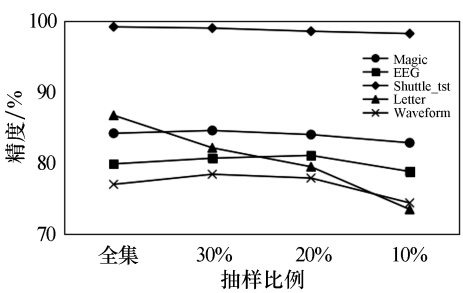

图 3　各抽样比例的精度

3.3　不同算法的约简效果

为了评价本文算法的分类精度，将本文算法得到的属性子集的分类精度与 Dependence 算法、BMLNRS 算法进行对比. BMLNRS 算法和本文算法的 δ 取 0.5σ，β 取 0.5，Dependence 算法则采用分类精度最好的结果. 用 3NN、SimpleCart、SMO、Bagging、JRip、RandomForest 算法计算每个属性子集的精度并取平均值，结果如表 6 所示.

表6 3种算法的属性约简精度

数据集	属性个数	Dependence算法			BMLNRS算法			本文算法		
		属性约简个数	全集精度	随机子集精度	属性约简个数	全集精度	随机子集精度	属性约简个数	全集精度	随机子集精度
Magic	10	8	84.68	84.04	8	84.46	84.01	8	84.49	83.91
EEG	14	12	84.82	80.45	9	78.71	79.94	10	79.88	80.19
Shuttle_tst	9	6	99.14	98.54	6	99.21	98.63	6	99.21	98.63
Letter	16	16	89.61	80.34	10	86.60	77.93	9	86.74	78.37
Waveform	40	8	74.49	74.51	8	77.00	76.90	6	75.49	75.54

从表6可以看出,本文算法的分类精度和BMLNRS算法基本相当. Dependence算法在EEG和Letter数据集上的精度优于本文算法,这是由于在这两个数据集上,Dependence算法约简后的属性个数多于本文算法,即保留的信息量多于本文算法.

4 结论

UCI数据集实验证明,本文提出的基于多次随机抽样的集成属性约简算法的空间占用比BMLNRS算法可减少2～3个数量级,且其约简精度和BMLNRS算法相当,所以本文方法在处理大规模数据时,具有更大的优势. 本文在生成约简子集时,仅考虑了一种属性评价标准,该评价标准可能会更偏好个别属性,因此今后将考虑综合多种评价标准,以进一步提高本文方法的鲁棒性.

参考文献:

[1] 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.

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

[3] CHEN H M, LI T R, CAI Y, et al. Parallel attribute reduction in dominance-based neighborhood rough set[J]. Information Sciences, 2016,373:351-368.

[4] LIN Y, LI J, LIN P, et al. Feature selection via neighborhood multigranulation fusion[J]. Knowledge-Based Systems, 2014,67(3):162-168.

[5] 鲍丽娜,丁世飞,许新征,等. 基于邻域粗糙集的极速学习机算法[J]. 济南大学学报(自然科学版),2015,29(5):367-371.

[6] LI X J, RAO F. Outlier detection using the information entropy of neighborhood rough sets[J]. Journal of Information & Computational Science, 2012,12(9):3339-3350.

[7] 沈林. 基于改进辨识矩阵的变精度邻域粗糙集属性约简[J]. 延边大学学报(自然科学版),2018,44(2):149-154.

[8] ZOU Q, ZENG J, CAO L, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification[J]. Neurocomputing, 2016,173(1):346-354.