基于非对称变邻域粗糙集模型的属性约简

惠景丽 潘 巍 吴康康 周晓英
(首都师范大学信息工程学院 北京 100048)
(首都师范大学高可靠嵌入式系统技术北京市工程研究中心 北京 100048)
(首都师范大学电子系统可靠性技术北京市重点实验室 北京 100048)

摘 要 在分析邻域粗糙集模型弊端的基础上,提出了非对称变邻域粗糙集模型,并以全局属性重要度为启发条件,构造了基于非对称变邻域粗糙集模型的属性约简的启发式算法。利用6个UCI标准数据集与现有算法进行了比较分析,结果表明,该模型不仅可以选择较少的属性个数,而且还能保持较高的分类能力。

关键词 邻域粗糙集,全局定邻域,非对称变邻域,全局属性重要度
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.059

Attribute Reduction Based on Asymmetric Variable Neighborhood Rough Set

HUI Jing-li PAN Wei WU Kang-kang ZHOU Xiao-ying
(College of Information Engineering, Capital Normal University, Beijing 100048, China)
(Beijing Engineering Research Center of High Reliable Embedded System, Capital Normal University, Beijing 100048, China)
(Beijing Key Laboratory of Electronic System Reliability Technology, Capital Normal University, Beijing 100048, China)

Abstract On the basis of analyzing the disadvantage of neighborhood rough set model, we proposed an asymmetric variable neighborhood rough set model and a new heuristic attribute reduction algorithm based on asymmetric variable neighborhood rough set. The heuristic condition is global attribute significance. Experimental results show that the number of attribute reduction and classification accuracy based on asymmetric variable neighborhood rough set model have better performance

Keywords Neighborhood rough set, Global neighborhood, Asymmetric variable neighborhood, Global attribute significance

1 引言

Pawlak教授在1982年提出的经典粗糙集理论(RST)[1]利用等价关系对论域进行等价类划分,并在此基础上定义了下近似(正域)和上近似算子来描述研究对象。但是,此模型只适合处理离散数据,对于连续数据,如果直接采用等价关系进行处理,将每一个连续的数据视为一个离散的数据值,则会出现过学习现象,从而影响数据的分类性能。通常可以有两种方法来解决这一问题:1)连续数据的离散化[2-4];2)直接扩展粗糙集模型[13-15],将其用于连续数据的处理中。离散化方法虽然相对简单,不需要建立新的模型,但是,采用离散化算法不可避免地带来了一定的信息损失[5-9]。

为了解决这一问题,Lin[10-12]在1998年借助于拓扑学中内点和闭包的概念提出了邻域系统,它可以有效处理连续数据。其后,Yao[13]、Wu[14]等对此模型的性质进行了进一步的研究和分析,并取得了一些进展。2008年,Hu[15,16]在前人研究的基础上提出了邻域粗糙粗模型,并将所提出的模型用于决策系统的属性约简和规则提取等现实问题的推理中。

目前,利用邻域粗糙集模型处理决策系统时[17],通常采用的是全局定邻域,即每个样本在不同的条件属性组合上采用的是同一个邻域值δ[16]。此模型是以参考样本为中心,左右选择一个相等的邻域值δ作为此样本的邻域范围。然而,通过实验[15,16]发现:1)δ值的微小变化都会对属性约简和规则提取产生很大影响;2)通过搜索算法得到最优的δ值,往往需要较高的时间复杂度,甚至最终得不到最优的δ值。这是因为数据分布多种多样,固定的邻域值以及对称的邻域结构在大多数情况下不能对它们进行精确的描述,分类结果会随着δ值的变化产生较大的波动,进而影响模型的稳定性。

综上,本文首先在邻域粗糙集模型的基础上,提出非对称变邻域粗糙集模型。它采用非对称结构的邻域形式,对不同的属性选择不同的邻域值,这种处理方式能够进一步提高对数据分布的描述能力。然后,在这种非对称变邻域的基础上提出了新的属性约简模型和算法,并采用UCI公开数据集验证属性约简算法的有效性。实验发现,与其他方法相比,该模

到稿日期:2014-08-20 返修日期:2015-01-06 本文受国家自然科学基金资助项目(61070049,61202027),国际科技合作项目(2012DFA11340),北京市自然科学基金资助项目(4122015),电子系统可靠性技术北京市重点实验室2012年阶梯计划项目(Z121101002812006)资助。
惠景丽(1985-),女,硕士生,主要研究方向为模式识别、信息融合,E-mail:huijingli085@163.com;潘 巍(1973-),女,博士,副教授,主要研究方向为模式识别、信息融合;吴康康(1989-),女,硕士生,主要研究方向为模式识别、信息融合;周晓英(1990-),女,硕士生,主要研究方向为模式识别、信息融合。

型不仅可以得到较少的条件属性个数，并且还能保证较高的识别率。此外，该模型能够根据数据分布特征，自适应地得到 δ 值，进一步提高了粗糙集的应用价值。

2 邻域粗糙集模型 NRSM

经典粗糙集模型(RSM)将信息系统定义为 $IS=\langle U,A\rangle$，其中，$U=\{x_1,x_2,\cdots,x_n\}$ 是非空对象集合；$A=C\cup D$，$C=\{c_1,c_2,\cdots c_m\}$ 为条件属性集，D 为决策属性。此模型是在离散空间中利用等价关系对 U 进行等价类划分，如果直接采用等价关系对连续数据进行处理，会出现过学习现象，影响数据的分类性能。目前，已有一些数据离散化方法使 RSM 能够处理连续型数据[2-4]。但是，对连续数据离散化不可避免地会引入更多的不精确的信息，最终导致错误的决策。为解决这一问题，Hu[15,16]在 Lin[10]邻域系统的基础上，将等价关系扩展为邻域关系，提出了邻域粗糙集模型。

2.1 NRS 的基本概念及性质

定义 1 给定一信息系统 $IS=\langle U,C\cup D\rangle$，$\forall x_i\in U$，$B\subseteq C$，$x_i$ 在属性空间 B 中的邻域 $\delta(x_i)$ 定义为：$\delta_B(x_i)=\{x_j\mid x_j\in U,\Delta^B(x_i,x_j)\leqslant\delta\}$，其中 C 是描述 U 的实数型特征集合，D 是决策属性，Δ 为距离尺度。此时该系统成为邻域决策系统 $NDT=\langle U,C\cup D,\delta\rangle$。

从定义 1 可以看出，NRSM 中的每个属性以及多个属性的集合都采用的是同一个邻域值 δ，即此邻域是一个全局定邻域值。这样的模型相对比较简单，但是从实验[15,16]可知，其训练出一个较优的定邻域值需要很高的时间复杂度。

定义 2 给定样本集 U，对于任意样本集 $X\subseteq U$，则 X 的下近似和上近似定义为：

$$\underline{NX}=\{x_i\mid\delta(x_i)\subseteq X,x_i\in U\}$$

$$\overline{NX}=\{x_i\mid\delta(x_i)\cap X\neq\varnothing,x_i\in U\}$$

定义 3 给定一邻域决策系统 $NDT=\langle U,C\cup D,\delta\rangle$，$\{X_1,X_2,\cdots,X_n\}$ 为论域 U 被决策属性 D 所分成的等价类集合，$\delta_B(x_i)$ 为样本 x_i 在条件属性 $B\subseteq C$ 下的邻域。条件属性 B 相对于决策属性 D 的上、下近似定义如下：

$$\overline{N_B}(D)=\bigcup_{i=1}^{n}\overline{N_B}X_i,\ \underline{N_B}(D)=\bigcup_{i=1}^{n}\underline{N_B}X_i$$

决策属性 D 的下近似也称为决策正域，记为 $POS_B(D)$，即 $POS_B(D)=\underline{N_B}(D)$。在某种程度上，正域可以反映决策属性在给定的条件属性空间中的确定性程度，因此，可以使用正域来粗略地确定某个条件属性相对于其他属性的重要程度。

定义 4 给定一邻域决策系统 $NDT=\langle U,C\cup D,\delta\rangle$，$\forall B\subseteq C$，则决策属性 D 对条件属性集 B 的依赖度为：$\gamma_B(D)=\dfrac{|\underline{N_B}(D)|}{|U|}$。

定义 5 给定一邻域决策系统 $NDT=\langle U,C\cup D,\delta\rangle$，$\forall B\subseteq C$，$\forall a\in C-B$，则 a 对 B 的重要度定义为：$SIG(a,B,D)=\gamma_{B\cup a}(D)-\gamma_B(D)$。

定理 1[15] 给定一邻域决策系统 $NDT=\langle U,C\cup D,\delta\rangle$，$B_1,B_2\subseteq C,B_1\subseteq B_2$，则 $POS_{B_1}(D)\subseteq POS_{B_2}(D)$。

此定理说明了相对正域 $POS_B(D)$ 是单调不减的，这对构建基于邻域粗糙集模型的属性相对约简算法至关重要。

定义 6 对邻域决策系统 $NDT=\langle U,C\cup D,\delta\rangle$，$B\subseteq C$，若 B 是一个相对约简(或者决策表的约简)，则需要满足以下两个条件：

(1)充分条件：$POS_B(D)=POS_C(D)$；

(2)必要条件：$POS_B(D)>POS_{B\setminus\{b\}}(D)$。

这两个条件反映了当条件属性增加时决策表的确定性结构，在属性约简中起到非常重要的作用。条件(1)是为了确保正域单调不减，即要求约简必须要保证决策表中所有条件属性与总表相比其分辨能力不变；条件(2)为了确保 B 中的每个属性都是必要的，在约简中不存在多余的属性，即每个属性都是相对独立的。

3 非对称变邻域粗糙集模型

在 NRSM 中，所选择的邻域是一个全局的对称定邻域 δ，它忽略了数据的局部分布差异，这使得 NRSM 的属性约简结果和分类性能不稳定。针对以上模型的不足，本文在邻域粗糙集模型的基础上，提出了一种非对称变邻域粗糙集模型。

数据分布的不均匀使得在各处的数据密度不同，即对于参考样本 x_i 的邻域的两个方向按照密度的差异存在着密部(DP)和疏部(SP)。因此，本文采用了非对称的邻域结构，在样本 x_i 的 DP 方向选择一个较小的邻域值 δ^D，而在 SP 方向需要选择一个较大的邻域值 δ^S。此时，信息系统 $AVN=\langle U,C\cup D,\delta^D(x_i),\delta^S(x_i)\rangle$ 为非对称变邻域信息系统。

定义 7 给定一非对称变邻域决策系统 $AVN=\langle U,C\cup D,\delta^D(x_i),\delta^S(x_i)\rangle$，$a\in C$，非对称变邻域粗糙集的单个属性邻域定义为(见图 1)：

$$\delta_a(x_i)=\{x_j\mid x_j\in U,\Delta^a(x_c(i),x_j)\leqslant\delta_a^*\}$$

其中，邻域值的半径为 $\delta_a^*=\dfrac{V_a^D(x_i)+V_a^s(x_i)}{2}$，邻域直径的中点值为

$$x_i^c(a)=\begin{cases}x_i(a)+\dfrac{V_a^s(x_i)-V_a^D(x_i)}{2}, & x_D<x_S\\ x_i-\dfrac{V_a^s(x_i)-V_a^D(x_i)}{2}, & \text{others}\end{cases}$$

$V_a^D(x_i)$ 为样本 x_i 在属性 a 上的密部邻域值，$V_a^s(x_i)$ 为疏部邻域值。

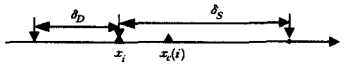

图 1 单个属性的邻域范围

为了使非对称邻域从一维向高维扩展变得更加直观，需要在不对称邻域和对称邻域之间建立映射，将非对称邻域用虚拟的中心点表示，即对应定义 7 中的 $x_i^c(a)$。

定义 8 给定一非对称变邻域决策系统 $AVN=\langle U,C\cup D,\delta^D(x_i),\delta^S(x_i)\rangle$，多个属性邻域的定义为(见图 2)：

$$AVN_B(x_i)=\{x_j\mid x_j\in U,\Delta^B(x_i^c,x_j)\leqslant\delta_B^*\}$$

其中，$B\subset U$，$B=\{B_k\mid k=1,2,\cdots,n\}$，$\delta_B^*=(\sum_{k=1}^{n}\delta_{B_k}^{*q})^{1/q}$，$q$ 为名可夫斯基系数，是一个非负整数。当 $q=2$ 时其为欧氏距离。

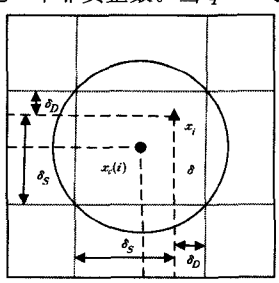

图 2 二维空间的邻域范围

设参考样本为 x_1，它在两个属性 a_1 和 a_2 上的值分别为 t_a 和 t_b。对于邻域粗糙集模型，样本 x_1 在属性 a_1 和 a_2 上的邻域范围分别为 $[t_a-\delta,t_a+\delta]$ 和 $[t_b-\delta,t_b+\delta]$，样本 x_1 在属

性集$\{a_1,a_2\}$上的邻域范围为以x_1为圆心、以δ为半径的圆，如图3(a)所示。而对于非对称变邻域粗糙集模型，样本x_1在属性a_1和a_2上的邻域范围分别为$[t_a-V_{a_1}^S(x_1),t_a+V_{a_1}^D(x_1)]$和$[t_b-V_{a_2}^S(x_1),t_b+V_{a_2}^D(x_1)]$，样本$x_1$在属性集$\{a_1,a_2\}$上的邻域范围为以$(x_i^s(a_1),x_i^s(a_2))$为圆心、以$\delta_{a_1\cup a_2}^*$为半径的圆，如图3(b)所示。

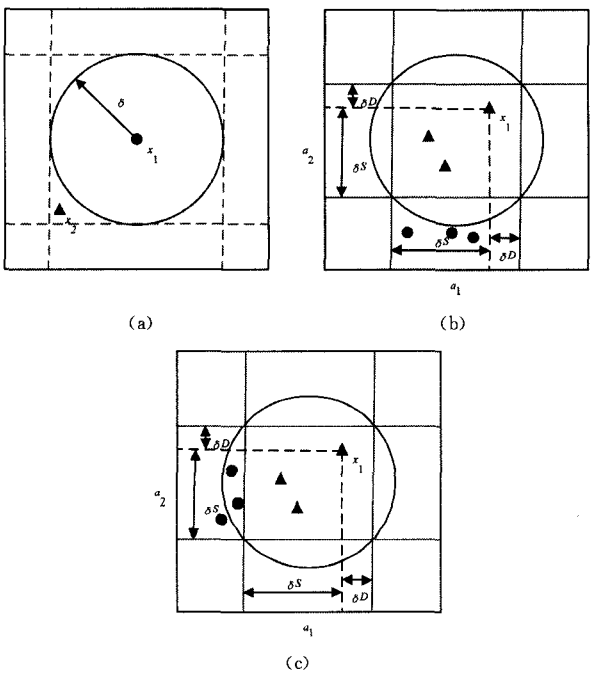

图3　在二维空间上不同邻域模型的邻域范围

根据定义1，邻域粗糙集模型的邻域范围为正方形的内切圆，因此，$\delta_{a_1\cup a_2}(x_1)\subseteq\delta_{a_1}(x_1)$，即多属性结合后，邻域粒度将减小，正域将不断地增加。例如，对属性a_1来说，如果样本x_1的邻域范围内所有样本同类，则a_1的正域必然包含样本x_1，并且不论x_1的邻域范围内的样本是否全为同类，属性组合(a_1,a_2)的邻域内(即圆形区域)的样本必然全为同类，即保证正域的不减性。不同地，在非对称变邻域粗糙集模型中，由于每个样本在每一维属性上邻域值都是不同的，在多维属性(比如二维属性)上，正域为交叉部分(即矩形区域)的外接圆，这就导致正域不再是单调不减的，而是可增可减的，因此正域的变化变得比较复杂。

(1)例如，在图3(b)中，不同类型的点代表不同类别的样本。样本x_1不属于属性a_1的正域，这是因为样本x_1在属性a_1的邻域范围内有其他类别的样本；但是在增加了属性a_2后，邻域变成了矩形区域的外接圆，而圆形区域的样本与样本x_1都是同类，所以x_1属于属性组合(a_1,a_2)的正域，此时，正域随着属性个数的增加而增加。

(2)例如图3(c)，样本x_1在属性a_1上的邻域内全是同类样本，所以x_1属于属性a_1的正域；但是在增加了属性a_2后，邻域内出现了不同类的样本，所以样本x_1就不属于属性组合(a_1,a_2)的正域，此时，正域就随着属性个数的增加而减小。

(3)当属性添加的顺序不同时，正域的变化也有可能是不同的。如图3(b)所示，样本x_1在属性a_2的邻域范围内的样本都是同类的，此时样本x_1属于属性a_2的正域；当增加了属性a_1后，区域内的样本同样是同类的，所以x_1仍然属于属性组合(a_1,a_2)的正域。此时，正域的大小没有变化。

综上所述，在非对称变邻域粗糙集模型中，正域并不是随属性的变化而单调变化的。这就导致基于邻域粗糙集模型的约简算法无法满足基于非对称变邻域粗糙集模型的属性约简算法的条件。

4　基于非对称变邻域粗糙集的前向贪心属性约简

属性约简是在不降低信息系统分类能力的前提下，在全体属性集中寻找一个最小的属性集，摒弃冗余的属性，从而降低对数据处理时的时间复杂度和空间复杂度。因此，属性约简被广泛应用于数据处理和知识发现中[18,19]，比如模式识别、机器学习和数据挖掘。

从第3节分析可以看出，在非对称变邻域粗糙集模型中，由于正域的非单调性，因此以定义6作为终止条件的前向约简算法已经无法得到较优的约简结果。因此，需要改进现有的属性约简算法。在基于非对称变邻域粗糙集模型的属性约简中不仅要更多地考虑正域的非单调性，还要考虑添加属性后对整个正域的影响。为了满足需要，非对称变邻域粗糙集的属性约简算法的终止条件定义如下。

定义9　对信息系统$IS=\langle U,C\cup D\rangle$，$B\subset C$，若$B$是一个相对约简(或者决策表的约简)，则需要满足以下两个条件：

(1)收敛条件：$\forall b\in C\backslash B, POS_B(D)\geqslant POS_{B\cup\{b\}}(D)$；

(2)不减条件：$POS_B(D)\geqslant POS_C(D)$。

定义9不仅能保证约简中不会出现冗余的属性(收敛条件)，还能使所得的约简有较高的分类能力(不减条件)。

在非对称变邻域粗糙集模型中，由于属性添加的顺序不同，正域的变化情况也有可能发生变化，因此要调整现有约简算法中的启发条件——属性重要度。下面给出以非对称变邻域粗糙集模型为基础的约简算法中的启发条件——全局属性重要度。

定义10　给定信息系统$IS=\langle U,C\cup D\rangle$，$a\in C-red$，属性$a$的全局重要度为

$$Sig(a,n)=\sum_{i=1}^{n}POS_{red\cup a}(D)$$

其中，n是迭代的次数，red是第$n-1$次迭代的约简。

定义10的叠加形式能很好地反映任一属性在条件属性不断增加时的整体变化情况。这是因为随着属性的添加，每个新添入的属性的重要度在当前的约简中是在不断发生变化的，而这种叠加形式能很好地体现该属性的整体情况。接下来，将给出基于非对称变邻域模型的前向贪心属性约简算法。在介绍约简算法前，首先确定样本在单个条件属性上的邻域。

本文主要根据数据自身的分布特点来确定邻域值，在分布比较密集的方向上取一个较小的邻域值，而在分布密度较稀疏的方向取一个相对较大的邻域值。具体算法如下：

令C为属性集，$C=\{C_1,C_2,\cdots,C_n\}$，样本集$X=\{x_1,x_2,\cdots,x_M\}(M<m)$。在一维空间中，“+”代表样本点右侧，“−”代表左侧。$V_{x_i}(a)$是样本x_i在属性a上的值。具体来讲，要确定任意样本x_i在某一属性a上的邻域，需要经过2个步骤：定位和扩展。

定位：确定样本的密部和疏部。在a上分别找到样本x_i左、右两侧的k个近邻；然后分别计算左、右两侧最远的第k个样本到样本x_i的距离d^-和d^+。距离值越大，则样本分布越稀疏，称为疏部(SP)；距离越小，则样本分布越密集，称为密部(DP)。例如，如果$d^->d^+$，则左侧为疏部，右侧为密部，

$\delta_D^i = d^+$，$\delta_S^i = d^-$；反之亦然。

扩展：根据样本的分布密度，按照不同的步进值对邻域值进行扩展。在实际应用中，中值准则的应用非常广泛。为了能够自适应地确定非对称邻域值的大小，这里同样采用中值准则来确定邻域值。

在这一步骤中，本文选择一个小步进 δ_D^i/κ 作为密部的邻域扩展步进值，选择一个较大的步进 δ_S^i/κ 作为疏部的步进值，如此不断扩展，直到满足中值准则。即不断检查 $|V_{x_i}(a) - mid(x_i)| \leqslant \varepsilon$ 是否成立，其中 $mid(x_i)$ 是位于区间 $[V_{x_i}(a) - \delta_D^i/\kappa, V_{x_i}(a) + \delta_S^i/k]$ 内样本集的中值，ε 为误差限。如果不等式成立，那么 $[V_{x_i}(a) - \delta_D^i/\kappa, V_{x_i}(a) + \delta_S^i/\kappa]$ 就作为样本 x_i 在属性 a 上的邻域。反之，则左右两侧邻域各增加一个各自的步进值，即区间变为 $[V_{x_i}(a) - 2 * \delta_D^i/\kappa, V_{x_i}(a) + 2 * \delta_S^i/\kappa]$，重复步骤直到不等式成立。此外，为了减小计算的复杂度，扩展的范围不超过 k 个数据，若均不能达到要求，则选择使 $|V_{x_i}(a) - mid(x_i)|$ 最小的邻域。至此，得到了样本在单属性上的邻域。

下面举例说明确定样本在单个条件属性上的邻域。

假设要得到样本 t 在属性 a 上的邻域，如图 4(a)所示，横轴为样本在属性 a 上的分布。(1)定位：首先找到 t 左、右两侧的 3 个近邻样本 (x_1, x_2, x_3) 和 $(x_4\ x_5, x_6)$，在这 6 个样本中距 t 最远的两个样本分别为 x_1 和 x_6。x_1 和 t 之间的距离为 d^-，x_6 和 t 之间的距离为 d^+，其中 $d^- > d^+$，所以左边为疏部，右边为密部。

(2)扩展：设 $\varepsilon = 0$，按照各自的步进值将邻域扩展后，新加入的样本为 x_4，如图 4(b)所示。显然 t 不是邻域的中值，所以需要将邻域扩展一倍，如图 4(c)所示。扩展后的邻域包含样本 x_3 和 x_4，从图中可以看出 t 是中值，所以 $\delta_D^i(a) = 4$，$\delta_S^i(a) = 1.7$。至此，可以根据定义 8 得到参考样本多个属性集合的邻域。

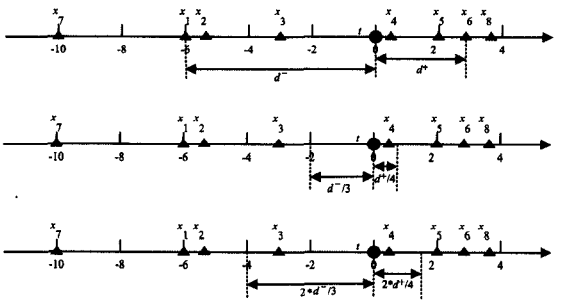

图 4 一维空间的非对称变邻域

算法 基于全局属性重要度的属性约简

Input：IS=〈U，C∪D〉

Output：约简 red

Step 1 red=∅

Step 2 ∀ For each $x_i \in U$

计算该样本对应条件属性上的非对称变邻域；//根据文中的介绍计算单个属性或多个属性集所对应的邻域范围，进而得到样本的非对称邻域

Step 3 for each $a_k \in C - red$

计算 $Sig(a_k, n) = \sum_{i=1}^{n} POS_{red \cup a_k}(D)$ //计算全局属性重要度，n 为迭代的次数

Step 4 if $\forall a_k \in C - red, POS_{red}(D) < POS_{red \cup \{a_k\}}(D)$

在下次选择属性时不考虑 a_k

Step 5 if $\forall a_k \in C - red, Sig(a_k, n) = Sig(a_k, n-1)$//算法收敛，输出约简 red

return red

elseif a_k, s. t. $\max(Sig(a_k, n))$

$red = red \cup a_k$//添加 a_k 到 red 中，如果 a_k 能使 $Sig(a_k, n)$ 最大

Step 6 go to Step 2

5 实验及分析

为了证明本文提出的模型和基于此模型的约简算法的有效性，本文使用 6 个 UCI 公开数据集对算法进行测试。此外，本文还采用了目前比较流行的约简算法与本文算法进行对比，并且采用 3 种分类器(BF Tree、CART Tree 和 C4.5)对不同的约简方法进行评估。本文采用这些分类器的识别率作为属性约简的评价标准是因为它们被广泛应用于不同类型的数据识别问题，而且分类思想和粗糙集近似。因此，使用这些分类器进行数据集的识别率判断更能够说明本文模型及算法的有效性。表 1 给出了所用数据集的一些描述，分别为数据集的属性个数、类别数目和样本数。从表 1 可以看出，所列的 6 个数据表的特征各异，几乎涵盖了大多数数据表的类型，数据表 Sonar、BCW 和 Wdbc 的属性个数较多，数据表 Glass 的决策类数目较多，而数据表 BCW 和 Wdbc 具有较多的样本数目。因此，对这些数据表的测试能更好地验证本文模型和约简算法的有效性和普适性。

表 1 数据集概况

Data	Attribute	Class	Sample
Glass	9	6	214
Wine	13	3	178
Sonar	60	2	208
BCW	30	2	569
Parkinsons	22	2	195
Wdbc	30	2	569

为了减小数据不同量纲对数据集带来的影响，在约简前首先需要将每个属性对应的属性值进行归一化，归一化所使用的公式为：$y = \frac{x - x_{min}}{x_{max} - x_{min}}$，其中，$x$ 是原始属性值，x_{min} 和 x_{max} 分别为原始值中的最小值和最大值，y 是归一化后的属性值。

表 2 不同约简算法得到的约简结果

Data	CFS	F2HARNRS	ours
Glass	3,2,4,6,7,1,8,5(8)	8,3,2,1,5,9,4(7)	4,3,1,8,7,2(6)√
Wine	1,2,3,4,5,6,7,10,11,12,13(11)	13,10,7,6,1(5)√	7,10,1,13,11,5,6(7)
Sonar	2,4,5,8,9,10,11,12,13,21,24,28,37,44,45,46,47,48,51,52,54,55,60(23)	58,51,55,1,6,8,35,45,20,30,26,17,11,39,22,34,9(17)	11,43,16,36,45,22,28,4,46(9)√
BCW	2,7,8,14,19,21,23,24,25,27,28(11)	23,28,2,29,5,16,25,9(8)	21,23,28,8,4,3,7(7)√
Parkinsons	15,18,19,20,22(5)√	19,16,1,18,20,3(6)	22,1,4,5,16,2,15,7(8)
Wdbc	4,7,8,14,19,21,22,23,27,28(10)	23,8,22,25,29,19(6)√	21,23,28,3,8,4,7,2,6,10(10)
Average	11.3	8.17	7.8√

表 2 给出了 3 种属性约简的算法：CFS、F2HARNRS

(Fast Forward Heterogeneous Attribute Reduction based on Neighborhood Rough Set)[16]和本文算法对数据表进行实验所得到的约简结果。其中,括号前的数字代表属性的序号,括号内的数字代表约简后属性的个数;"√"代表3种方法中约简后属性个数最少的组合;最后一行代表6个数据集经3种不同的约简方法后的平均个数。

从表2可以看出,3种基于不同模型的属性约简算法都能有效减少条件属性的个数。但是本文算法在3个数据集上取得了最少的属性约简个数,并且使得属性个数较多的数据集Sonar约简后的属性个数减少得更加明显。从约简属性个数的值来看,基于本文方法的约简属性的个数也最少。正如前面介绍,本文方法是根据数据自身的分布特点,在每个样本对应的属性上都能自适应地找到一个非对称的邻域值,从而更能发挥不同属性的分类能力,使得每个属性的相对正域较其他方法有所增大。虽然基于本文模型的正域具有非单调性,但是它在一定程度上减少了达到最大正域的属性个数,从而使本文方法得到较少的约简属性个数。

一个较优的属性约简结果不仅要有较少的约简属性个数,还需要有较好的分类能力。为了评估本文模型所得到的约简在分类精度上的性能,本文采用BF、CART和C4.5这3种分类算法分别对3种模型所得到的约简进行分类精度的测试,并将结果汇总到表3—表5中。其中第二列为原表的识别率;"√"代表其中最优的组合,即识别率最高的一组;"↑"代表该结果较原表识别率有所提高;"↓"代表该结果较原表识别率有所下降;"—"代表该结果与原表识别率持平;最后一行表示所有数据集的平均识别率。

表3 用CART测试不同属性约简结果的识别率(%)

Data	Raw	CFS	F2HARNRS	ours
Glass	70.56	72.43√	69.63↓	71.5↑
Wine	89.33	89.33 —	89.33 —	89.89√
Sonar	71.15	75.00↑	76.92√	75.96↑
BCW	92.44	93.32↑	92.90↑	93.85√
Parkinsons	89.23	87.18↓	83.59↓	90.26√
Wdbc	93.15	93.85↑	92.80↓	94.38√
Average	84.31	85.18↑	84.19↓	85.97√

表4 用BF测试不同属性约简结果的识别率(%)

Data	Raw	CFS	F2HARNRS	ours
Glass	64.49	72.43√	71.03↑	69.63↑
Wine	89.89	89.89√	89.33↓	89.33↓
Sonar	71.63	75.96↑	77.88√	77.88√
BCW	92.79	93.15↑	92.44↓	94.73√
Parkinsons	90.26	88.20↓	86.15↓	90.26√
Wdbc	92.97	92.79↓	93.50↑	94.2√
Average	83.67	85.4√	85.05↑	86↑

表5 用C4.5测试不同属性约简结果的识别率(%)

Data	Raw	CFS	F2HARNRS	ours
Glass	65.89	70.09↑	67.76↑	71.96√
Wine	93.82	93.82 —	94.94√	94.38↑
Sonar	71.15	73.08↑	76.92↑	78.85√
BCW	95.08	93.50↓	93.67√	93.5↓
Parkinsons	90.77	89.74↓	88.20↓	91.28√
Wdbc	93.15	93.32↑	94.20↑	94.55√
Average	84.98	95.59↑	85.95↑	87.42√

从表3—表5的结果可以看出,用3种分类方法测试的结果存在一定的差异,这主要是由于分类器自身分类算法的差异造成的。虽然3种测试结果不同,但是,从同种分类器的比较来看,本文方法总能得到较好的结果。即使没有得到最优的结果(比如表3中数据集Sonar,基于本文算法得到的识别率为75.96%,低于基于F2HARNRS算法的76.92%),基于本文算法得到的属性约简个数(9)也远少于基于F2HARNRS算法的约简个数(17),并且基于本文算法的识别率仍然比原表识别率(71.15%)高。

总的来说,基于本文模型的约简方法在3种分类测试方法中都能得到最优的识别结果,即使没有取得最优的结果,识别结果也接近最好的结果;而且从表2—表5的平均值结果可以看到,基于本文模型不仅可以得到最少的属性约简个数,并且识别率也最高。

结束语 本文介绍了邻域粗糙集的基本概念,针对邻域粗糙集的缺点,提出了一种新的非对称变邻域粗糙集模型。与HU提出的邻域粗糙集模型相比,本模型用非对称邻域和变邻域代替了邻域粗糙集模型中的对称邻域和全局定邻域,从而使邻域值能更加客观准确地反映数据自身的分布特征,最终得到较好的属性约简结果。此外,本文模型能够根据数据自身的分布特点自适应地找到合适的邻域值,而HU的模型只能通过多次实验试探地找到合适的邻域值。最后,采用6个UCI公开数据集验证了本文所提出模型及约简算法的有效性。

参考文献

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

[2] 赵军,王国胤,吴中福,等. 基于粗集理论的数据离散化方法[J]. 小型微型计算机系统, 2004, 25(01): 60-64
Zhao Jun, Wang Guo-yin, Wu Zhong-fu, et al. Method of Data Discretization Based on Rough Set Theory [J]. Mini-micro Systems, 2004, 25(01): 60-64

[3] 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 Hao-zhong, Niu Dong-xiao. Discretization of Continuous Attributes in Rough Set Theory Based on Information Entropy [J]. Chinese Journal of Computers, 2005, 28(9): 1570-1574

[4] 陈果. 基于遗传算法的决策表连续属性离散化方法[J]. 仪器仪表学, 2007, 28(9): 1700-1705
Chen Guo. Discretization method of continuous attributes in decision table based on genetic algorithm [J]. Chinese Journal of Scientific Instrument, 2007, 28(9): 1700-1705

[5] Jensen R, Shen Q. Semantics-Preserving dimensionality reduction: Rough and fuzzy-rough-based approaches[J]. IEEE Trans. on Knowledge and Data Engineering, 2004, 16(12): 1457-1471

[6] Wong Tzu-Tsung. A hybrid discretization method for naive Bayesian classifiers [J]. Pattern Recognition, 2012, 45(6): 2321-2325

[7] Augasta M G, Kathirvalavakumar T. A new discretization algorithm based on range coefficient of dispersion and skewness for neural networks classifier [J]. Applied Soft Computing, 2012, 12(2): 619-625

[8] Li Min, Deng Shao-bo, Feng Sheng-zhong, et al. An effective discretization based on Class-Attribute Coherence Maximization [J]. Pattern Recognition Letters, 2011, 32(15): 1962-1973

[9] Ferreira A J, Figueiredo M A T. An unsupervised approach to

feature discretization and selection [J]. Pattern Recognition, 2012,45(9):3048-3060

[10] Lin T,Granular Y. Computing on binary relations I:Data mining and neighborhood systems[C]// Skoworn A,Polkowshi L,eds. Proc. of the Rough Sets in Knowledge Discovery. Physica-Verlag,1998:107-121

[11] Lin Tsau-young. Encyclopedia on complexity of systems science [M]. Berlin:Springer,2009:4339-4355

[12] Yang Xi-bei, Li Xin-zhe, Lin Tsau-young. First GrC model neighborhood systems: the most general rough set models [C]// 2009 IEEE International Conference on Granular Computing. Beijing,China:IEEE,2009:691-695

[13] Yao Y Y. Relational interpretation of neighborhood operators and rough set approximation operators[J]. Information Sciences, 1998,111(198):239-259

[14] Wu W Z,Zhang W X. Neighborhood operator systems and approximations[J]. Information Sciences,2002,144(1-4):201-217

[15] 胡清华,于达仁,谢宗霞. 基于邻域粒化和粗糙集逼近的数值属性约简[J]. 软件学报,2008,19(3):640-649
Hu Qing-hua,Yu Da-ren,Xie Zong-xia. Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation [J]. Journal of Software,2008,19(3):640-649

[16] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Science, 2008,178(18):3577-3594

[17] 王丽娟,吴陈,杨习贝,等. 邻域系统粗糙集和覆盖粗糙集[J]. 计算机科学,2013,40(1):221-224
Wang Li-juan,Wu Chen,Yang Xi-bei,et al. Neighborhood System Based Rough Set and Covering Based Rough Set[J]. Computer Science,2013,40(1):221-224

[18] 张颖淳,苏伯洪,曹娟. 基于粗糙集的属性约简在数据挖掘中的应用研究[J]. 计算机科学,2013,40(8):223-226
Zhang Ying-chun,Su Bo-hong,Cao Juan. Study on Application of Attributive Reduction Based on Rough Sets in Data Mining [J]. Compute Science,2013,40(8):223-226

[19] 李智玲,胡彧. 改进的属性约简算法在数据挖掘中的应用研究[J]. 计算机技术与发展,2012,22(10):47-50
Li Zhi-ling,Hu Yu. Application Research of Improved Attribute Reduction Algorithm in Data Mining[J]. Computer Technology and Development,2012,22(10):47-50

(上接第 278 页)

和 4.45,*Rank2* 比 Baseline 平均出现相关文档个数高出近 55%,明显提高了反馈文档的质量。

结束语 为了尽可能地将满足用户信息需求的文档查询出来,减少查询语义缺失,提出了一种基于最大边缘相关的伪相关反馈方法。通过对初检结果进行两次重定序,既保证了与查询相关文档的优先返回,又可以有效避免伪相关反馈中的查询主题偏移问题。本文的相似度算法仅选取了基于向量空间的 cosine 系数来计算文档间相似度,进一步的工作是选取更合适的相似度计算方法,提高算法运行速度。

参 考 文 献

[1] Collins-Thompson K. Reducing the risk of query expansion via robust constrained optimization[C]// Proceedings of the 18th ACM conference on Information and Knowledge Management. ACM,2009:837-846

[2] Raman K,Udupa R,Bhattacharya P,et al. On improving pseudo-relevance feedback using pseudo-irrelevant documents[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2010:573-576

[3] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval[C]// Proceedings of the 10th international conference on Information and Knowledge Management. ACM,2001:403-410

[4] He B, Ounis I. Studying query expansion effectiveness[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2009:611-619

[5] Vargas S,Santos R L T,Macdonald C,et al. Selecting effective expansion terms for diversity[C]// Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. 2013:69-76

[6] Clarke C L A,Kolla M,Cormack G V,et al. Novelty and diversity in information retrieval evaluation[C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM,2008:659-666

[7] Song R,Luo Z,Nie J Y,et al. Identification of ambiguous queries in web search[J]. Information Processing and Management, 2009,45:216-229

[8] Amati G,Carpineto C,Romano G. Query difficulty,robustness, and selective application of query expansion[M]// Advances in information retrieval. Springer Berlin Heidelberg,2004:127-137

[9] Parapar J, Presedo-Quindimil M A, Barreiro á. Score distributions for Pseudo Relevance Feedback[J]. Information Sciences, 2014,273:171-181

[10] Teevan J,Dumais S T,Horvitz E. Characterizing the value of personalizing search[C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,2007:757-758

[11] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summarics [C]// Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,1998:335-336

[12] Cronen-Townsend S,Zhou Y,Croft W B. Predicting query performance[C]// Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,2002:299-306

[13] Shah C, Croft W B. Evaluating high accuracy retrieval techniques[C]// Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,2004:2-9

[14] Karimzadehgan M, Zhai C. A learning approach to optimizing exploration-exploitation tradeoff in relevance feedback[J]. Information retrieval,2013,16(3):307-330