文章编号:1000－2367(2017)05－0068－06 DOI:10.16366/j.cnki.1000－2367.2017.05.012

邻域决策一致性的属性约简方法研究

李智远[1],杨习贝[2a,3],徐苏平[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.计算机科学与工程学院;
b.数理学院,江苏 镇江 212003;3. 南京理工大学 经济管理学院,南京 210094)

摘 要:基于邻域决策错误率的属性约简可以在删除冗余属性的同时,提升邻域分类器的留一验证分类精度.但这种约简方式并未充分考虑邻域分类结果在约简前后的差异.为解决这一问题,借助联合分布矩阵,提出了邻域决策一致性的概念,构建了邻域决策一致性与邻域分类精度的调和平均值,并将其作为约简求解的度量准则.在 12 个 UCI 数据集上的实验结果表明,所提出的新约简不仅能够有效地提升邻域分类器的决策一致性,而且在多数情况下能够进一步提高邻域分类器的留一验证分类精度.

关键词:邻域分类器;邻域决策错误率;邻域决策一致性;约简

中图分类号:TP18 **文献标志码**:A

邻域粗糙集[1－2]是经典粗糙集[3]的一种重要拓展模型.邻域粗糙集使用距离的方法来度量样本之间的相关程度,并在此基础上利用半径来直接决定邻域信息粒的大小.近年来,邻域粗糙集方法因其具有简洁直观的表达方式、对复杂数据具有较强的适应性、易于实现增量式计算等诸多优点得到了众多学者的广泛关注[4－10].

类似于其他粗糙数据分析方法,属性约简在邻域粗糙集的理论与应用研究中也占据着核心地位.从粗糙集本身描述不确定性的视角来看,可以以近似质量、条件熵、粗糙近似分布等度量指标来定义基于邻域粗糙集的属性约简.然而更重要的是,在邻域粗糙集理论中,还可以从分类学习的视角来研究属性约简问题.例如,邻域决策错误率[11]概念的提出,就是在使用邻域分类器[12]进行分类学习的基础上,讨论如何找到较小的属性子集,来降低邻域决策发生错误的程度,从而提高邻域分类器的分类精度.

基于邻域决策错误率的属性约简虽然具有目标清晰、易推广等优点,但不难发现这种约简的目的仅仅是降低使用邻域分类器进行决策分析的错误程度,并没有将约简前后邻域分类结果的差异性充分地体现出来.单纯地追求邻域决策错误率的降低,其本质是希望原始数据中被错误分类的那些样本,在约简后能够尽量多地被正确分类,但其忽视了约简之后有另外一种不一致性,即在原始数据中被正确分类的样本,在约简后有可能就会被错误分类,显然,在分类学习中这是不希望被看到的.为尽量降低由约简带来的这种不一致性,有必要在约简过程中找到这些不一致的样本,并尽量地减少这些样本带来的影响.借助集成学习中的联合分布矩阵[13]概念,可以找到这些不一致的样本.此外,为了在求解约简时,保证追求不一致性降低的同时,邻域决策错误率也能够有所降低,故本文提出了将一致性与精度的调和平均作为新的约简标准.

本文主要内容安排如下:第 1 节简要介绍邻域决策错误率的基本概念;第 2 节在邻域决策错误率约简的基础上,提出了邻域决策一致性约简的概念;第 3 节进行实验对比分析;第 4 节总结全文.

收稿日期:2016-12-10;**修回日期**:2017-05-10.

基金项目:国家自然科学基金(61572242;61503160;61502211;61471182);江苏省高校哲学社会科学基金(2015SJD769);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,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),研究方向为粗糙集理论、粒计算、机器学习,E-mail:zhenjiangyangxibei@163.com.

1　邻域分类器与邻域决策错误率

在粗糙集理论中，一个决策系统可以被描述为二元组 $\Omega=(U,A_T\cup\{d\})$，其中 U 是所有样本所构成的集合，A_T 是所有条件属性集合，d 是决策属性，$A_T\cap\{d\}=\varnothing$. $\forall a\in A_T$，V_a 是条件属性 a 的值域，V_d 是决策属性 d 的值域. 因本文专注于分类问题，故 $V_d=\{1,2,\cdots,m\}$ 中是一些离散属性值，记录了所有样本的类别标记，$d(x)\in V_d$ 是样本 x 的类别标记. $U/IND(\{d\})=\{X_1,X_2,\cdots,X_m\}$ 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类别范畴，为简化问题描述起见，等价类的下标就是该等价类对应的类别标记.

邻域粗糙集采用距离这一概念度量决策系统中任意两个样本之间的相似程度. $\forall x,y\in U$，$\delta_{A_T}(x,y)$ 用来表示样本 x 与 y 在属性集合 A_T 上的距离度量，这一度量可以使用欧式距离，余弦距离等不同的距离计算方法. 在此基础上，给定一半径 $\sigma\in\mathbf{R}$，$\delta_{A_T}(x)=\{y\in U:\delta_{A_T}(x,y)\leq\sigma\}$ 表示利用半径 σ 所构建的 x 的邻域，其语义解释是所有与 x 在半径 σ 下被认为是相似的对象所构成的集合.

利用邻域信息粒，除了可以构建邻域粗糙集以外，亦可设计邻域分类器进行分类学习研究，算法 1 给出了邻域分类器[12]的详细流程. 邻域分类器与传统的近邻分类器，其分类思想类似，但不同之处在于近邻分类器是指定邻居的个数，而邻域分类器则是通过半径来圈定邻居，因此，不同的样本可能会产生不同的邻居个数.

算法 1：邻域分类器

输入：决策系统 Ω，待预测样本 y，邻域半径 σ.

输出：y 的预测类别标记 $\rho_{A_T}(y)$.

步骤 1　$\forall x\in U$，计算 $\delta_{A_T}(y,x)$；

步骤 2　计算 $\delta_{A_T}(y)$；

步骤 3　$\forall X_i\in U/IND(\{d\})$，计算 $\Pr(X_i,\delta_{A_T}(y))=\dfrac{|\delta_{A_T}(y)\cap X_i|}{|\delta_{A_T}(y)|}$；

步骤 4　$X_j=\arg\max\{\Pr(X_i,\delta_{A_T}(y)):\forall X_i\in U/IND(\{d\})\}$；

步骤 5　$\rho_{A_T}(y)=j$，输出 $\rho_{A_T}(y)$.

利用邻域分类器，Hu 等[11]进一步给出了决策系统中邻域决策错误率的概念，用以构建属性约简的度量准则. 邻域决策错误率的计算思想是每次将 U 中的一个样本作为待测试样本，剩余样本作为训练样本，利用邻域分类器预测待测试样本的类别，直到 U 中的每一个样本都被作为测试样本使用过一次，最终得到邻域分类器的错误率，作为邻域决策错误率. 显然，这是一种留一验证方法，以下给出邻域决策错误率的形式化定义.

定义 1　令 Ω 为一决策系统，Ω 中的邻域决策错误率(NDER，简记为 N)为

$$N_{A_T}=\frac{|\{x\in U:\rho_{A_T}(x)\neq d(x)\}|}{|U|}.\tag{1}$$

从分类学习的视角来看，$1-N_{A_T}$ 就是邻域分类器的留一验证分类精度.

2　属性约简

利用邻域决策错误率的概念，Hu 等[11]进一步地给出了相应的属性约简描述如定义 2 所示.

定义 2　令 Ω 为一决策系统，$\forall A\subseteq A_T$，A 被称为 Ω 的一个邻域决策错误率约简(NDERR，简记为 N^R)当且仅当 $N_A\leq N_{A_T}$ 且对于任意的 $B\subset A$，都有 $N_B>N_{A_T}$.

由定义 2 可以看出 Ω 中的邻域决策错误率约简是一个使得 Ω 中的邻域决策错误率能够被降低的最小属性子集，同时由定义 1 不难发现，利用邻域决策错误率约简，邻域分类器的留一验证分类精度能够有所提升.

虽然利用定义 2 可以得到比原始数据更高的留一验证分类精度，但这种约简策略并未充分考虑约简前

后邻域分类器的分类结果,换言之,有可能会出现约简前后分类结果不一致的情形,例如在原始数据中被正确分类的样本,利用约简进行分类时,有可能会被错误分类. 鉴于此,需要在约简中,进一步考虑邻域分类器的分类结果. $\forall A \subseteq A_T$,表 1 描述了一个联合分布矩阵,表示利用属性集合 A 和属性集合 A_T 所产生分类结果的分布情况.

表 1 邻域分类器的联合分布矩阵

	$\rho_A(x)=d(x)$	$\rho_A(x)\neq d(x)$
$\rho_{A_T}(x)=d(x)$	a	b
$\rho_{A_T}(x)\neq d(x)$	c	d

根据表 1,可以得知: a 表示在 U 中,利用 A_T 和 A 都能够被正确分类的样本数目,b 表示在 U 中,利用 A_T 能够被正确分类而利用 A 却被错误分类的样本数目,c 表示在 U 中,利用 A_T 被错误分类而利用 A 却被正确分类的样本数目,d 表示在 U 中,利用 A_T 和 A 都被错误分类的样本数目. 在机器学习研究中,联合分布矩阵可以用来度量两个基分类器的分类差异性[13],而本文则利用该矩阵计算不同属性集合下利用同一种分类器所得到分类结果的差异性.

定义 3 令 Ω 为一决策系统,$\forall A \subseteq A_T$,A 相对于 A_T 的分类差异性(DISAG,简记为 D)记为

$$D(A,A_T)=\frac{b+c}{a+b+c+d}=\frac{b+c}{|U|}. \tag{2}$$

显然,在定义 3 中,有 $0 \leq D(A,A_T) \leq 1$ 成立,$1-D(A,A_T)$ 被称为 A 相对于 A_T 的邻域决策一致性度量. $1-D(A,A_T)=1$ 意味着最高的一致性,即由 A 所得到的分类结果与原始属性集合 A_T 所得到的分类结果是完全一致的,对于在 A_T 下被正确 / 错误分类的那些样本,在 A 下也能够被正确 / 错误分类;$1-D(A,A_T)=0$ 意味着最低的一致性,即由 A 所得到的分类结果与原始属性集合 A_T 所得到的分类结果是完全不一致的,对于在 A_T 下被正确分类的那些样本,在 A 下却被错误分类了,而在 A_T 下被错误分类的那些样本,在 A 下却被正确分类了.

在属性约简的过程中,如果单纯地追求决策一致性较高,将会致使没有任何属性被认为是冗余的,这也会直接造成邻域分类精度无法进一步得到提升. 此外,追求分类一致性较高,意味着联合分布矩阵中 $b+c$ 的值将会降低,c 是那些利用 A_T 被错误分类而利用 A 却被正确分类的样本数目,若 c 值降低,则就意味着有可能会损失一定的分类精度. 所以为了在一致性与分类精度之间取得平衡,本文利用邻域决策错误率和邻域决策一致性,构建了一种调和平均(HAV,简记为 H) 形如:

$$\frac{1}{H_A}=\frac{1}{2}\cdot\left(\frac{1}{1-N_A}+\frac{1}{1-D(A,A_T)}\right). \tag{3}$$

定义 4 令 Ω 为一决策系统,$\forall A \subseteq A_T$,A 被称为 Ω 的一个邻域决策一致性约简(NDAR) 当且仅当 $H_A \geq H_{A_T}$ 且对于任意的 $B \subset A$,都有 $H_B < H_{A_T}$.

定义 4 所示的调和平均约简,其目的是寻求使得邻域分类精度和邻域决策一致性的调和平均能够被提升的最小属性子集.

以下给出一个启发式算法用以求解定义 4 所示的约简.

算法 2:邻域决策一致性约简

输入:决策系统 Ω.

输出:邻域决策一致性约简 A.

步骤 1 $A=\varnothing$;

步骤 2 $\forall a \in A_T$,计算 $H_{\{a\}}$;

步骤 3 $A=A \cup \arg\max\{H_{\{a\}}:\forall a \in A_T\}$;

步骤 4 若 $H_A \geq H_{A_T}$,则转步骤 6,否则转步骤 5;

步骤 5 当 $H_A < H_{A_T}$ 时执行以下循环:

(1) $\forall b \in A_T - A$,计算属性 a 的重要度 $H_{A \cup \{b\}}$;

(2) $A=A \cup \arg\max\{H_{A\cup\{b\}}:\forall b \in A_T - A\}$;

步骤 6　输出 A.

3　实验分析

为了验证本文提出约简定义的有效性，选取了 12 组 UCI 数据集进行实验分析，数据的基本描述如表 2 所列. 实验环境为 PC 机，双核 2.60 GHz CPU，16 GB 内存，Windows10 操作系统，MATLAB R2010a 实验平台.

表 2　数据集描述

ID	数据集	样本个数	属性个数	类别个数
1	Cardiotocography	2126	21	10
2	Dermatology	366	34	6
3	Glass Identification	214	9	6
4	Libras Movement	360	90	15
5	Molecular Biology	106	57	2
6	Parkinson Multiple Sound Recording	1208	26	2
7	QSAR biodegradation	1055	41	2
8	Seeds	210	7	3
9	Sonar	208	60	2
10	Statlog (Vehicle Silhouettes)	846	18	4
11	Wisconsin Diagnostic Breast Cancer	569	30	2
12	Wine	178	13	3

在本组实验中，选取了 10 个不同的半径 σ，值分别是 0.05，0.1，…，0.5. 在这 10 个半径取值下，图 1 给出了原始数据的分类精度，利用邻域决策错误率约简(NDERR)所得到的留一验证分类精度、邻域决策一致性度量，以及利用邻域决策一致性约简(NDAR)所得到的留一验证分类精度、邻域决策一致性度量. 图 1 所示的实验结果均为十折交叉验证求解约简后得到的分类精度与一致性度量的平均值. 该实验的主要目的是为了对比利用两种不同约简所得到的邻域决策一致性，以及对比利用原始数据及两种不同约简所得到的分类精度.

通过观察图 1，不难得到如下结论.

(1)在 10 个不同的半径 σ 下，利用邻域决策一致性约简所得到的决策一致性度量都明显高于利用邻域决策错误率约简所得到的决策一致性度量，这说明本文所提出的邻域决策一致性约简可以有效地提升邻域分类结果的一致性.

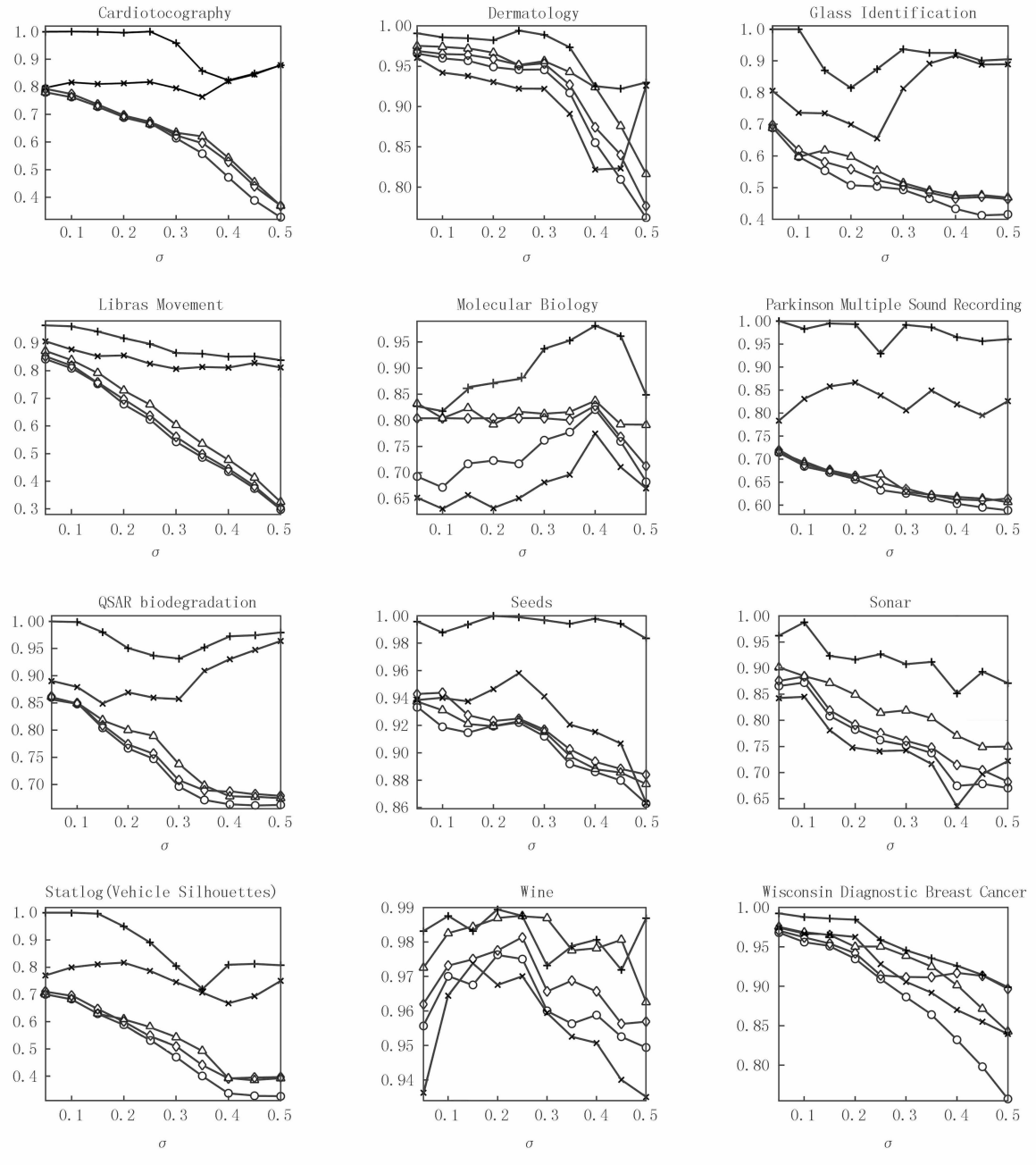
(2)在 10 个不同的半径 σ 下，利用邻域决策错误率约简和邻域决策一致性约简所得到的分类精度都明显高于原始数据中的邻域分类精度，这说明文中涉及的两种约简方法都能够在不同程度上降低邻域错误分类率，但同时值得注意的是，利用邻域决策一致性约简所得到的分类精度在大多数数据集及半径 σ 的取值上，都明显高于利用邻域决策错误率约简所得到的分类精度(仅在 Seeds 数据上，邻域决策一致性约简的分类精度略显不足，但与邻域决策错误率约简的分类精度差距很小)，这说明利用本文所提出的调和平均度量求解约简，可以在提升决策一致性的同时，进一步提升邻域分类器的分类精度.

4　结　论

在分析了邻域决策错误率约简未能充分考虑约简前后分类结果的不一致性这一不足之处，提出了邻域决策一致性约简的概念，这种新的约简将邻域决策一致性度量与邻域决策分类精度的调和平均作为度量指标，在公开数据集上的实验结果表明，新约简能够在有效地提升分类结果的一致性，同时也能保证邻域分类器的留一验证分类精度有进一步的提升空间. 在本文工作的基础上，将就以下工作进行深入探讨：

(1)由代价敏感问题带来的邻域分类不一致性问题；

(2)为进一步提升约简的时间效率,需寻求更高效的快速求解算法.



注:⊖ 原始精度;◇ NDERR的精度;△ NDAR的精度;✱ NDERR的一致性;+ NDAR的一致性.

图1 不同约简下的分类精度及一致性

参 考 文 献

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

[2] 杨习贝,杨静宇. 邻域系统粗糙集模型[J]. 南京理工大学学报(自然科学版),2012,36(2):291-295.

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

[4] Hu Q H,Zhu P F,Yang Y B,et al. Large-margin nearest neighbor classifiers via sample weight learning[J]. Neurocomputing,2011,74(4):656-660.

[5] He Q,Xie Z X,Hu Q H,et al. Neighborhood based sample and feature selection for SVM classification learning[J]. Neurocomputing,2011,74(10):1585-1594.

[6] 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.
[7] 张维，苗夺谦，高灿，等. 邻域粗糙协同分类模型[J]. 计算机研究与发展，2014，51(8)：1181-1820.
[8] 段洁，胡清华，张灵均，等. 基于邻域粗糙集的多标记分类特征选择算法[J]. 计算机研究与发展，2015，52(1)：56-65.
[9] Lin Y J，IIu Q II，Liu J II，et al. Multi-label feature selection based on neighborhood mutual information[J]. Applied Soft Computing，2016，38：244-256.
[10] 吕康，魏培文，张辉. 基于粒计算的融合性贴近度方法[J]. 河南师范大学学报(自然科学版)，2015，43(5)：153-158.
[11] IIu Q II，Pedrycz W，Yu D R，et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Transactions on Systems，Man，and Cybernetics，Part B，2010，40(1)：137-150.
[12] IIu Q II，Yu D R，Xie Z X. Neighborhood classifiers[J]. Expert Systems with Applications，2008，34(2)：866-876.
[13] Kuncheva L，Whitaker C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J]. Machine Learning，2003，51(2)：181-207.

Attribute Reduction Approach to Neighborhood Decision Agreement

Li Zhiyuan[1]，Yang Xibei[2a,3]，Xu Suping[2a]，Chen Xiangjian[2a]，Wang Pingxin[2b]

(1. Kewen College，Jiangsu Normal University，Xuzhou 221116，China；2. a. School of Computer Science and Engineering；
b. School of Mathematics and Physics，Jiangsu University of Science and Technology，Zhenjiang 212003，China；
3. School of Economics & Management，Nanjing University of Science and Technology，Nanjing 210094，China)

Abstract：Attribute reduction based on neighborhood decision error rate can improve the leave-one-out classification accuracy of neighborhood classifier via deleting redundant attributes. Nevertheless，such approach does not fully take the difference between classification results before and after reduct into account. To solve such problem，from the viewpoint of joint distribution matrix，the neighborhood decision agreement is proposed and a new criterion for attribute reduction is constructed，which is the harmonic mean of neighborhood decision agreement and neighborhood classification accuracy. The experimental results on 12 UCI data sets show that the new criterion based reduct can not only improve the decision agreement of neighborhood classifier，but also the leave-one-out classification accuracy of neighborhood classifier will also be increased in most cases.

Keywords：neighborhood classifier；neighborhood decision error rate；neighborhood decision agreement；reduct

[责任编辑　陈留院]