文章编号: 1001-0920(2018)07-1207-08 DOI: 10.13195/j.kzyjc.2017.0361

混合值不完备系统的双邻域粗糙集分类方法

黄恒秋[1†], 曾 玲[2], 黎利辉[1]

(1. 广西民族师范学院 数学与计算机科学学院，广西 崇左 532200；
2. 桂林电子科技大学 数学与计算科学学院，广西 桂林 541004)

摘 要: 针对混合值不完备系统,提出一种基于双邻域粗糙集模型的分类方法. 首先,定义一个新的不确定距离度量函数 —— 联系度距离函数,进而建立基于联系度距离函数的双邻域粗糙集模型;然后,基于所建立的模型讨论该模型的属性约简算法,并给出基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法;最后,通过多个UCI 数据集进行实证分析,结果表明所提出的分类算法是客观有效的,特别是在缺失值较多的情况下,其优势更加明显.

关键词: 混合值不完备系统；双邻域粗糙集；联系度距离；分类

中图分类号: TP18 **文献标志码:** A

Double-neighborhood rough set classification method in incomplete decision system with hybrid value

HUANG Heng-qiu[1†], *ZENG Ling*[2], *LI Li-hui*[1]

(1. School of Mathematics and Computer Science, Guangxi Normal University for Nationalities, Chongzuo 532200, China; 2. School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: In order to process the incomplete decision system with hybrid value, a classification method based on the double-neighborhood rough set model is proposed in this paper. Firstly, an uncertain distance function — Connection degree distance function is defined, and the double-neighborhood rough set model based on connection degree distance function is constructed. Then, based on the constructed model, an attribute reduction algorithm is discussed, and a classification algorithm based on attribute reduction and covering reduction is provided. Finally, some experiments are carried out on UCI data sets. The experiments results show that the proposed classification algorithm is objective and effective, and it is more effective when the missing value is abounded.

Keywords: incomplete decision system with hybrid value；double-neighborhood rough set；connection degree distance；classification

0 引 言

粗糙集模型[1]由Pawlak于1982年提出,是处理不精确、不一致、不完备数据的智能信息处理技术. 传统粗糙集模型只适合处理完备的符号型数据,针对现实应用中广泛存在的数值型数据则不能直接处理[2]. Lin[3]通过拓展等价关系,首次将邻域关系引入粗糙集模型中,该模型将空间中点的邻域作为基本信息粒子,并用来描述空间中的其他概念,但是没有针对现实问题进行应用分析. Hu等[4]针对现实中的数据系统,采用距离函数定义邻域,给出了能够同时处理数值型与符号型数据的邻域粗糙集模型.

邻域粗糙集模型的主要研究内容包括属性约简和规则学习[5],它们构成了邻域粗糙集模型分类技术的核心. 基于邻域粗糙集构造分类器主要有3种方式: 1) 构建基于邻域粗糙集模型的混合分类器. 首先采用邻域粗糙集模型进行属性约简,然后对约简后的系统采用诸如支持向量机、朴素贝叶斯、决策树等主流分类方法进行分类[4,6-11]. 2) 构建具有决策能力的邻域决策粗糙集模型. 如文献[12-13]给出了基于决策理论的邻域决策粗糙集模型、三支邻域决策粗糙集模型以及相关的分类技术. 3) 构建基于邻域粗糙集规则学习的分类技术. 首先采用邻域粗糙集模型

收稿日期: 2017-03-29；**修回日期:** 2017-07-10.
基金项目: 广西重点培育学科(应用数学)建设项目(SXYB2014005, SXYB2016001)；国家民委科研项目(14GSZ015).
责任编委: 刘民.
作者简介: 黄恒秋(1983−), 男, 讲师, 从事数据挖掘、粗糙集理论及其应用的研究；曾玲(1963−), 女, 教授, 从事决策分析、不确定理论及其应用等研究.
†通讯作者. E-mail: hengqiu0417@163.com

进行属性约简或覆盖约简并提取规则,然后基于提取的规则给出分类算法. 如文献[14]首次给出了基于邻域粗糙集模型的规则分类算法;文献[15]将邻域粗糙集规则学习分类技术应用于医疗诊断;文献[16]给出了基于覆盖约简的规则学习分类算法;文献[17]给出了基于覆盖约简的双邻域规则学习分类算法;文献[5]讨论了基于规则学习分类技术的邻域粒度选择问题及自适应选择算法;文献[9]针对存在多个有效约简的问题,给出了基于规则集成学习的分类算法.

基于邻域粗糙集规则学习的分类技术充分利用了邻域局部信息,且规则易于理解和实现[10],目前已成功应用于各种分类任务[9],但是针对广泛存在的混合值不完备系统的讨论较少. 本文将同时存在符号属性、数值属性及缺失值的系统称为混合值不完备系统. 目前,处理该类系统的方法主要有两类:一类是选择具备处理缺失数据能力的距离函数,比如文献[4,5,16-18]采用的混合距离(HEOM)[19]. 但是,正如文献[20]指出,HEOM将缺失属性值距离取作最大值,如果属性较多且样本属性值缺失比例较大时,会造成系统信息失真. 因此,建议修改为可用的比较属性取值全相等时,缺失属性值距离取0,否则取1,即介于最悲观和最乐观之间. 文献[18]则是取最乐观的情形. 文献[21-24]虽然对HEOM距离作了改进,但是均为针对属性加权和完备数据进行讨论. 另一类是将邻域关系拓展为邻域容差关系. 比如文献[25-26]将缺失值看作与任何属性值都相等的邻域容差关系和拓展邻域关系;文献[27]则通过定义样本对的3种联系度,即同一度、对立度和差异度,并设置对立度为0,同一度大于给定的阈值,给出了邻域联系度容差关系.

本文将讨论混合值不完备系统的双邻域粗糙集规则学习分类框架. 首先,为了避免缺失值距离取极端值,同时又能充分利用同一度、对立度和差异度在不确定对象相似性度量方面的优势,定义一种新的不确定距离度量函数——联系度距离函数;然后,借鉴文献[17]能够有效处理噪音数据的双邻域近似思想,构建基于联系度距离的双邻域粗糙集模型,并给出该模型的属性约简算法;最后,以联系度距离函数作为距离度量,给出基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法.

为了检验本文方法的有效性,取7个UCI测试集进行了实验分析. 首先通过与HEOM缺失值距离取最乐观、最悲观两种情况获得的分类精度进行对比;然后,通过与文献[24-26]中拓展邻域关系、邻域容差关系、HEOM距离度量相关分类技术取得的分类精度进行了比较,结果表明本文提出的联系度距离函数和分类方法均取得了较佳的效果.

1 邻域系统相关概念

1.1 几种邻域关系

定义1[4] 给定决策系统$I=(U,A\bigcup D,V,f)$,$B\subseteq A$,定义属性空间B上的δ-邻域关系为

$$\delta_B(x_i)=\{x_j|x_j\in U,\ \Delta(x_i,x_j)\leqslant\delta\}.$$

其中:$\delta\geqslant 0$,Δ是一个距离函数.

定义2[25] 设$I=(U,A\bigcup D,V,f)$为不完备系统,$B\subseteq A$,$B=B_C\bigcup B_N$,B_C为符号属性,B_N为数值属性,$\delta\geqslant 0$. 由B确定的邻域容差关系定义为

$$\begin{aligned}&\mathrm{NT}_B^{\delta}(x)=\\&\{(x,y)\in U^2|\forall\, a\in B(a(x)=*\bigvee a(y)=\\&*\bigvee((a\in B_C\to\Delta_a(x,y)=0)\bigwedge\\&(a\in B_N\to\Delta_a(x,y)\leqslant\delta)))\}.\end{aligned}$$

其中:“$\to$”表示需满足的条件,“$*$”表示缺失值.

定义3[27] 给定混合值不完备系统$I=(U,A\bigcup D,V,f)$,$|A|=N$,Δ为绝对值距离函数,$(x_i,x_j)\in U^2$为集对. 记ε为属性值相容的邻域半径,设$M=\{a\in A|\Delta_a(x_i,x_j)\leqslant\varepsilon\}$为集对取值在相容邻域范围之内的属性集;$H=\{a\in A|\Delta_a(x_i,x_j)>\varepsilon\}$为集对取值在相容邻域范围之外的属性集;$G=\{a\in A|f(x_i,a)=*\bigvee f(x_j,a)=*\}$为集对取值不明确的属性集. 记$m=|M|/N$,$g=|G|/N$,$h=|H|/N$,则集对$(x_i,x_j)$的邻域联系度可以表示为

$$\mu(x_i,x_j)=m+gi^*+hj^*.$$

其中:m、g、h记作同一度、差异度和对立度;i^*、j^*为差异度和对立度标记,起到与同一度区别的作用. 取$0\leqslant t\leqslant 1$,$B\subseteq A$,定义B上的邻域联系度容差关系为

$$\begin{aligned}\delta_B^t(x_i)=\{(x_i,x_j)\in U^2|&\mu(x_i,x_j)=m+gi^*,\\&m+g=1,\ m\geqslant t\}.\end{aligned}$$

从定义3可以看出,两个样本同一度大于或等于给定阈值t,且对立度为0,可以归为同一类样本.

定义4[26] 给定混合值不完备系统$I=(U,A\bigcup D,V,f)$,$B\subseteq A$,$P_B(x)=\{a\in B|f(x,a)\neq *\}$,$\delta\geqslant 0$,则属性子空间$B$上的拓展邻域关系为

$$\begin{aligned}V_B(x,y)=\{(x,y)\in U^2|&(\Delta(x,y)\leqslant\delta)\bigvee\\&((f(x,a)=f(y,a)\bigvee f(x,a)=\\&*\bigvee f(y,a)=*)\bigwedge(\mu(x,y)\geqslant\alpha))\}.\end{aligned}$$

其中

$$\mu(x,y)=|P_B(x)\bigcap P_B(y)|/|P_B(x)|,\ \alpha\in[0,1].$$

定义5[17]　记样本x_i的Tri-分割邻域为

$$O_\beta(x_i)=\{x_j|\Delta(x_i,x_j)\leqslant\delta^\beta,\ (x_i,y_j)\in U^2\}.$$

其中: Δ为距离函数; β为邻域异质度,即邻域样本中与x_i不同类别的样本数与邻域样本总数之比; δ^β是异质度为β时的邻域半径. 则Tri-分割邻域的下近似邻域和上近似邻域分别定义为

$$O_*(x_i)=\{x_j|\Delta(x_i,x_j)\leqslant\delta^{\beta=0},x_j\in U\},$$
$$O^*(x_i)=\{x_j|\Delta(x_i,x_j)\leqslant\delta^{\beta=r},x_j\in U\}.$$

Tri-分割邻域的基本思想是通过邻域异质度来控制邻域近似精度,下近似邻域要求$\beta=0$,上近似邻域要求$\beta=r$,其中$r\in(0,1)$为给定阈值. 邻域半径的计算则是基于类间距实现的,文献[17]以x_i到达最近异类样本的距离与到达最近同类样本的距离之差作为下近似邻域半径,以使邻域异质度达到r的最远同类样本的距离作为上近似邻域半径. 通过双邻域近似处理技术,能够有效地处理噪音数据,提高分类精度.

从定义5可以看出,如果系统是不一致的,即属性取值完全相同或者相容而类别不同,则下近似邻域异质为0的要求不成立. 本文将在后面对其要求进行放宽,以适用不一致系统.

1.2　邻域系统中的距离度量

1) 闵氏距离[4](Minkowsky distance).

$$\Delta_p(x_1,x_2)=\Big(\sum_{i=1}^{m}|f(x_1,a_i)-f(x_2,a_i)|^p\Big)^{1/p},$$

其中m为向量维度. $p=1$时,为闵可夫斯基距离(Manhattan distance); $p=2$时,为常见的欧氏距离; $p=\infty$时,为切比雪夫距离(Chebychev distance).

2) 混合距离[20].

$$\mathrm{HEOM}(x,y)=\sqrt{\sum_{i=1}^{m}w_{a_i}\times d_{a_i}^2(x_{a_i},y_{a_i})}.$$

其中

$$d_{a_i}(x_{a_i},y_{a_i})=\begin{cases}1,\ x\text{ or }y\ \text{在}a_i\text{取值缺失};\\ \mathrm{overlap}_{a_i}(x,y),\ \text{符号属性};\\ \mathrm{rn\_diff}_{a_i}(x,y),\ \text{数值属性}.\end{cases}$$

$$\mathrm{overlap}_{a_i}(x,y)=\begin{cases}0,\ x_{a_i}=y_{a_i};\\ 1,\ x_{a_i}\neq y_{a_i};\end{cases}$$

$$\mathrm{rn\_diff}_{a_i}(x,y)=\frac{|x_{a_i}-y_{a_i}|}{\max(a_i)-\min(a_i)}.$$

2　基于联系度距离的双邻域粗糙集模型

2.1　联系度距离函数

从1.2节中混合距离(HEOM)的定义可以看出,如果样本x和y在属性a_i上的取值存在缺失,则缺失值距离由极端值1代替,文献[18]则是以另一个极端值0来代替. 文献[25]给出的邻域容差关系本质上与文献[18]相同. 文献[27]的邻域联系度容差关系要求对立度为0,而同一度通过阈值t去控制,要求严苛且阈值不好控制. 文献[26]给出的拓展邻域关系实质上是一种限制邻域容差关系,它对数据的完备程度进行了限制. 从定义3可以看出,同一度反映了两个样本的相同或者相容部分,最理想情况为1. 因此,将两个样本的同一度与最理想情况作比较,可获得其同一度的差异,而对立度和差异度本身就反映了两个对象之间的差异,将它们的差异通过加权的方式计算出来,即为联系度距离.

定义6　给定样本x,y的邻域联系度$\mu(x,y)=m+gi^*+hj^*$,则它们的联系度距离定义为

$$\mathrm{CDD}(x,y)=\sqrt{w_1\times(1-m)^2+w_2\times g^2+w_3\times h^2},$$

其中w_1, w_2, w_3为同一度、差异度和对立度的惩罚系数,且要求$w_1+w_2+w_3=1$.

从定义6可以看出,该距离函数继承了同一度、对立度和差异度在度量不确定样本相似性方面的优势,且利用的信息更加全面,避免了相关联系度阈值的选择问题,也避免了对缺失属性值人为干预填充或者取极端值的情形.

关于惩罚系数w_1,w_2,w_3. 事实上,两个样本属性取值不相同或者相异,最能够反映样本之间的差异,体现为对立度,因此惩罚系数应该最大; 差异度是由于样本属性值缺失造成的,缺失值有可能与比较样本相异,其惩罚系数次之; 同一度反映比较样本明确不相异部分,惩罚系数应该最小.

若取相容邻域半径$\varepsilon=0.1$,样本$x_1=(0.1,*,*,a,b,*)$, $x_2=(0.2,0.1,0.2,a,a,a)$, $x_3=(0.1,0.2,0.6,*,b,b)$,则样本对(x_1,x_2)和(x_3,x_2)的联系度分别为$\mu(x_1,x_2)=0.333+0.5i^*+0.167j^*$, $\mu(x_3,x_2)=0.333+0.167i^*+0.5j^*$. 如果不乘以惩罚系数,联系度距离相同,则无法区分. 若取惩罚系数$w_1=0.1,w_2=0.2,w_3=0.7$,则$\mathrm{CDD}(x_1,x_2)=0.338$, $\mathrm{CDD}(x_3,x_2)=0.474$. 事实上,样本对$(x_3,x_2)$具有3个相异的属性值,而$(x_1,x_2)$只有一个. 直观上看,其计算结果也是合理的.

为了更好地说明联系度距离在度量不确定样本相似性方面的优势,下面给出联系度距离与HEOM距离取两种极端情况时的比较计算例子(缺失值距离取最大值1,记为HEOMA; 缺失值距离取最小值0,记为HEOMB,本文下同,惩罚系数取值同上),具体的

计算数据见表1.

表1 一个混合值不完备决策系统

U	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	0.1	*	*	a	*	*	Y
x_2	0.1	0.15	*	b	b	*	Y
x_3	0.2	0.1	0.2	a	a	a	Y
x_4	0.2	*	0.5	b	*	*	N
x_5	0.1	0.2	0.6	b	b	b	N
x_6	0.2	0.15	0.5	*	b	b	N

例1 以各样本到x_6的距离为例.

联系度距离: 取相容邻域半径$\varepsilon = 0.1$, 则样本x_1, x_2, x_3, x_4, x_5到x_6的距离值分别为0.46、0.27、0.47、0.37、0.09, 各样本到x_6的距离大小排序结果为$x_3 > x_1 > x_4 > x_2 > x_5$.

HEOMA: 样本x_1, x_2, x_3, x_4, x_5与x_6的距离值分别为2.24、1.73、1.76、2、1.01, 各样本到x_6的距离大小排序结果为$x_1 > x_4 > x_3 > x_2 > x_5$.

HEOMB: 样本x_1, x_2, x_3, x_4, x_5与x_6的距离值分别为0.1、0.1、1.45、0、0.15, 各样本到x_6的距离大小排序结果为$x_3 > x_5 > x_1 = x_2 > x_4$.

直观上看: x_5和x_2与x_6相似性最大, 它们没有取值相异的属性值, 且完备性相对较高, 联系度距离和HOEMA距离排序结果相同; x_4和x_1与x_6虽然没有取值相异的属性值, 但是数据缺失比较严重, 由于HEOMA、HEOMB对缺失属性值分别取最悲观和最乐观值, 排序结果呈现了两个极端, 而联系度距离综合考虑了属性取值的相似性、相异性和完备程度, 其排序结果介于中间位置; x_3属性取值完备度最好, 但是与x_6有3个属性值相异, 应是最不相似的样本, HEOMB距离和联系度距离度量合理, 且排序结果相同.

基于以上分析可以看出, 联系度距离度量是有效的, 而且更加符合客观事实.

2.2 基于联系度距离的双邻域粗糙集模型

定义7 给定混合值不完备系统$I = (U, A \bigcup D, V, f)$, $B \subseteq A$. 记属性子空间B上关于样本x_i的β-划分邻域为

$$\delta_\beta(x_i) = \{x_j | \mathrm{CDD}_B(x_i, x_j) < \delta_{x_i}^{\beta},\ x_j \in U\}.$$

其中: CDD为联系度距离函数; $\beta \in [0, 1)$为邻域异质度, 即邻域样本中与x_i不同类别的样本数与邻域总样本数之比; $\delta_{x_i}^{\beta}$为异质度是β时样本x_i的邻域半径. 则x_i基于β-划分邻域的下近似邻域和上近似邻域分别定义为

$$\delta_*(x_i) = \{x_j | \mathrm{CDD}_B(x_i, x_j) < \delta_{x_i}^{\beta=\mathrm{nh}}, x_j \in U\},$$
$$\delta^*(x_i) = \{x_j | \mathrm{CDD}_B(x_i, x_j) < \delta_{x_i}^{\beta=r}, x_j \in U\}.$$

其中: 下近似邻域半径$\delta_{x_i}^{\beta=\mathrm{nh}} = \mathrm{CDD}(x_i, \mathrm{NM}(x_i)) - (\mathrm{CDD}(x_i, \mathrm{NH}(x_i)) + \eta)$, 上近似邻域半径$\delta_{x_i}^{\beta=r} = \mathrm{CDD}(x_i, \mathrm{FM}(x_i))$, $\mathrm{NM}(x_i)$为距离x_i的最近异类样本, $\mathrm{NH}(x_i)$为距离x_i的最近同类样本, $\mathrm{FM}(x_i)$为使样本x_i的邻域异质度达到r且邻域样本数量最大的样本, $r \in (0, 1)$; η为控制参数, 文中取$\eta = 0.000\,1$, 主要是为了防止最近同类样本与最近异类样本距离相同导致半径为0的情况. 这里$\beta = \mathrm{nh}$为样本x_i的下近似邻域异质度, 如果系统是一致的, 则$\mathrm{nh} = 0$, 否则$\mathrm{nh} \neq 0$.

定义8 给定论域U, 称$N_* = \{\delta_*(x_i) | x_i \in U\}$和$N^* = \{\delta^*(x_i) | x_i \in U\}$分别为下近似邻域粒度集合和上近似邻域粒度集合, 并称$\langle U, N_* \rangle$和$\langle U, N^* \rangle$为下近似邻域空间和上近似邻域空间.

定义9 对于任意的$X \subseteq U$, X的下近似、上近似分别定义为

$$N_*X = \{x_i | \delta_*(x_i) \subseteq X,\ x_i \in U\},$$
$$N^*X = \{x_i | \delta^*(x_i) \bigcap X \neq \phi,\ x_i \in U\}.$$

定义10 称(N_*X, N^*X)为X的双邻域粗糙集.

3 基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法

3.1 属性约简

定义11 给定混合值不完备系统$I = (U, A \bigcup D, V, f)$, $B \subseteq A$, 记β_i为x_i的下近似邻域异质度, 则系统下近似异质度和下近似同质度分别定义为

$$\beta_I = \Big(\sum_{i=1}^{|U|} \beta_i\Big) \Big/ |U|,$$
$$\gamma_I = 1 - \beta_I.$$

定义12 给定混合值不完备系统$I = (U, A \bigcup D, V, f)$, $B \subseteq A$, 属性$a \notin B$的重要度定义为

$$\mathrm{SIG}(a) = \gamma_I(B \bigcup a) - \gamma_I(B).$$

下面给出基于系统下近似同质度不变的属性约简算法.

算法1 基于系统下近似同质度不变的属性约简算法.

输入: $I = (U, A \bigcup D, V, f)$;

输出: 属性约简red.

Step 1: $\phi \rightarrow$red

Step 2: for each $a_i \in A -$red

计算$\mathrm{SIG}(a_i)$, end

Step 3: 选择a_k, 使得$\mathrm{SIG}(a_k) = \max(\mathrm{SIG}(a_i))$

Step 4: if $\mathrm{SIG}(a_k) > 0$

red $\bigcup a_k \rightarrow$ red, go to step 2

else return red, end.

3.2 基于属性约简、覆盖约简的双邻域粗糙集规则学习算法

定义13[16] 设$C = \{\delta(x_1), \delta(x_2), \cdots, \delta(x_n)\}$构成了论域$U$的逐点覆盖,称$\langle U, C\rangle$为一个邻域覆盖空间,$\langle U, C, D\rangle$为一个邻域覆盖决策系统.

定义14[16] $\langle U, C, D\rangle$为一个邻域覆盖决策系统,X_i为某一个决策类,如果存在$\delta(x_i') \in C$,使得$\delta(x_i') \subseteq \delta(x_i) \subseteq X_i$,则称$\delta(x_i')$相对于$X_i$是一致可约的,否则称$\delta(x_i')$是相对一致不可约的.

定义15[16] 给定$\langle U, C, D\rangle$,如果对于任意的决策类X_i,都不存在$\delta(x_i') \in C$,使得$\delta(x_i') \subseteq \delta(x_i) \subseteq X_i$,则称$\langle U, C, D\rangle$是相对不可约的,否则称$\langle U, C, D\rangle$是相对可约的.

定义16[16] $\langle U, C, D\rangle$是一个邻域覆盖决策系统,$C' \subseteq C$是从C中去除冗余覆盖元所得到的一个覆盖,且$\langle U, C', D\rangle$是相对不可约的,称C'是C的一个D相对约简.

算法2 基于属性约简、覆盖约简的双邻域粗糙集规则学习算法.

输入: $I = (U, A \bigcup D, V, f)$;

输出:基于属性约简、覆盖约简的下近似规则集合R_1和上近似规则集合R_2.

Step 1: 对于$I = (U, A \bigcup D, V, f)$,采用算法1,获得约简后的系统$I = (U, \text{red} \bigcup D, V, f)$.

Step 2: 对于任意的$x_i \in U$,构造双邻域近似集合$\{\delta_*(x_i), \delta^*(x_i)\}$,计算$\delta_{x_i}^{\beta=\text{nh}}$和$\delta_{x_i}^{\beta=r}$,并设$C_* = \bigcup\{\delta_*(x_i)\}$和$C^* = \bigcup\{\delta^*(x_i)\}$.

Step 3: 初始化$\phi \to R_1$和$\phi \to R_2$.

Step 4: 执行以下操作,获得R_1.

if $C_* = \phi$,则输出R_1

else 记$k = \max\limits_i\{n_i | n_i = |\delta_*(x_i)|, \delta_*(x_i) \in C_*, |\cdot|$表示集合元素个数$\}$,则$R_1 \bigcup (x_k, \delta_{x_k}^{\beta=\text{nh}}, d_k) \to R_1$,其中$d_k$表示$x_k$的类别

if $\exists \delta_*(x_p) \in C_*$使得$\delta_*(x_p) \subseteq \delta_*(x_k)$,则$\phi \to \delta_*(x_p)$

end, $\phi \to \delta_*(x_k)$

end

Step 5: 执行以下操作,获得R_2.

if $C^* = \phi$,则输出R_2

else 记$k = \max\limits_i\{n_i | n_i = |\delta^*(x_i)|, \delta^*(x_i) \in C^*, |\cdot|$表示集合元素个数$\}$,则$R_2 \bigcup (x_k, \delta_{x_k}^{\beta=r}, d_k) \to R_2$,其中$d_k$表示$x_k$的类别

if $\exists \delta^*(x_p) \in C^*$使得$\delta^*(x_p) \subseteq \delta^*(x_k)$,则$\phi \to \delta^*(x_p)$

end, $\phi \to \delta^*(x_k)$

end

3.3 基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法

由算法2获得了基于属性约简、覆盖约简的下近似规则集合R_1和上近似规则集合R_2. 下面给出规则学习分类算法.

算法3 基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法.

输入:测试集Test $= \{x_1, x_2, \cdots, x_m\}$和$R_1, R_2$;

输出:测试集的分类结果.

Step 1: 对于每个$x_i \in$ Test,分别计算其到下近似规则集$(x_j, \delta_{x_j}^{\beta=\text{nh}}, d_j)$和上近似规则$(x_t, \delta_{x_t}^{\beta=r}, d_t)$的联系度距离CDD$(x_i, x_j)$和CDD$(x_i, x_t)$. 其中: $j = 1, 2, \cdots, |R_1|, t = 1, 2, \cdots, |R_2|$.

Step 2:根据以下条件对x_i进行判别:

if $\exists k$使得CDD$(x_i, x_k) \leqslant \delta_{x_k}^{\beta=\text{nh}}$,则判定$x_i$的类别为$d_k (1 \leqslant k \leqslant |R_1|)$.

else if $\exists l$使得CDD$(x_i, x_l) \leqslant \delta_{x_l}^{\beta=r}$,将满足条件的所有$(x_l, \delta_{x_l}^{\beta=r}, d_l)$加入到规则候选集$O^c$中. 记$O^c$中到$x_i$距离最小的规则为$(x_k, \delta_{x_k}^{\beta=r}, d_k)$,则判定$x_i$的类别为$d_k (1 \leqslant l \leqslant |R_2|)$.

else令$\Delta(x_k) = \min\{\text{CDD}(x_i, x_j) - \delta_{x_j}^{\beta=nh}\}$,则判定$x_i$的类别为$d_k (1 \leqslant j \leqslant |R_1|)$.

4 实验分析

4.1 联系度距离函数有效性实验分析

从http://archive.ics.uci.edu/ml/下载7个UCI数据集,除Heart为完备数据集外,其他数据集均为不完备数据集,具体信息见表2.

表2 UCI数据集描述

数据集名称	样本数	类别数	属性数	数值属性数	符号属性数
Credit	690	2	15	6	9
Vote	435	2	16	0	16
Wpbc	198	2	33	33	0
Heart	270	2	13	5	8
Hepatitis	155	2	19	6	13
Soybean	683	19	35	0	35
Breast-Cancer	699	2	9	0	9

实验采用Matlab 2011B进行编程,数值属性值采用极差法全部规范化为[0, 1]之间,实验的分类精度均为10次交叉检验获得的平均精度. 双邻域粗糙集模型中,下近似邻域的半径根据定义7中的公式计算,而上近似邻域半径的计算公式中异质度取$r = 0.05$, HEOM距离公式中的权重均取1. 实验设计步骤如下.

首先,基于覆盖约简的双邻域粗糙集规则学习分类算法(距离函数分别采用联系度距离、HEOMA、HEOMB),对7个UCI数据集进行分类.

然后,对Heart数据集作随机缺失测试,缺失值增加比例依次为5%、10%、15%、20%、25%、30%,采用同样的分类算法和距离函数,对6种随机缺失测试数据集进行分类.

最后,针对以上两种情况,以分类精度来评价距离函数的优劣,并分析其原因.

1) 距离函数采用联系度距离进行分类算法实验分析. 关于惩罚系数的选择,文中2.1节从样本的相似性度量、各个联系度的重要程度方面进行了讨论,这里进一步从分类效果方面进行实验分析. 以混合属性的Hepatitis数据集为例,分原始数据(属性值缺失比例为5.67%)和增加5%随机缺失比例(不覆盖原缺失值,达到10.67%)两种情况做分析. 固定$w_1 = 0.1$, $\varepsilon = 0.15$,观察w_2, w_3的变化对分类效果的影响. 最终实验结果显示,在两种情况下,当w_3取0.7时均能够获得最优的分类精度. 因此,本文中取惩罚系数$w_1 = 0.1$, $w_2 = 0.2$, $w_3 = 0.7$是合理的;关于相容邻域半径的选择,对于不同数据集,其属性取值相容程度是不同的,但是因为数值属性都规范化到[0, 1]之间,所以取ε介于[0.05, 0.3](间隔取0.05),并通过实验从中选择最优分类精度作为比较是合理的. 同时,也取其平均值作为参考指标. 具体实验结果见表3中的第2、3列. 需要说明的是,以上ε取值同样适用于符号属性.

表3 联系度距离、HEOMA、HEOMB分类精度比较

数据集	CDD	CDD(avg)	HEOMA	HEOMB	缺失比例/%
Credit	0.839	0.828	0.843	0.833	0.65
Vote	**0.921**	0.921	0.928	0.439	**5.63**
Wpbc	0.747	0.731	0.768	0.763	0.06
Heart	0.837	0.815	0.822	0.822	0.00
Hepatitis	**0.847**	0.837	0.807	0.787	**5.67**
Soybean	**0.860**	0.860	0.838	0.478	**9.78**
Breast-Cancer	0.948	0.948	0.949	0.939	0.25

2) 距离函数采用HEOMA和HEOMB进行分类算法实验分析,其结果见表3的第4、5列. 通过对比可以看出:缺失属性值比例不到1%的4个数据集,其分类精度差别不大;缺失比例在5.63% ~ 9.78%的3个数据集,联系度距离要稍优于HEOMA距离,而HEOMB距离则变得非常糟糕.

3) 对Heart数据集作随机缺失测试结果如图1所示. 从图1可以看出,联系度距离在缺失比例较大情况下,其分类精度均优于HEOMA、HEOMB.

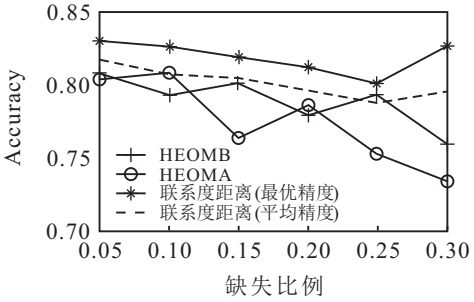

图1 Heart随机缺失比例测试结果

综上可以看出,在缺失值比例较小的情况下,联系度距离、HEOMA、HEOMB三种距离度量的分类效果差别不大. 随着缺失值比例增大,联系度距离度量的分类效果要优于HEOMA、HEOMB. 事实上,缺失值比例较大时,取极端值代替不仅容易造成系统信息失真,而且样本间的相似性也遭到破坏,导致分类效果变差. 联系度距离没有对缺失值作任何处理,保障了数据集的客观真实性,还充分利用了样本间的同一度、对立度和差异度信息,避免了其相似性遭受破坏,从而分类效果更佳.

4.2 基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法实验分析

4.1节中分析了联系度距离函数的有效性. 这里距离函数选择联系度距离,对基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法进行实验分析.

首先,采用本文给出的算法1对表2中的6个不完备数据集进行属性约简.

其次,利用本文给出的算法3进行分类,获得其平均精度和最优精度.

再次,从文献[24-26]中选择基于HEOM距离度量、邻域容差关系、拓展邻域关系相关分类方法获得的结果与本文结果(平均精度、最优精度)进行对比,见图2(星号曲线为约简精度,圆圈曲线为原始数据精度). 图2展示了6个数据集的分类效果,纵轴表示分类精度,横轴表示对应数据集采用的分类方法(用序号表示),详细说明如下.

Credit、Wpbc: 图2中星号曲线横轴对应的序号1~5依次表示文献[25]中基于(FSCE、SetCover、PR、IFSPA-IPR、SFFSNTCE)属性约简的朴素贝叶斯分类算法;6~9依次表示文献[26]中基于拓展邻域关系属性约简的(决策树、支持向量机、邻域粗糙集、K-最近邻)分类算法;10、11表示本文基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法,前者表示平均精度,后者表示最优精度. 图2中圆圈曲线横轴对应的序号1~6依次表示朴素贝叶斯分类算法、决策树分类算法、支持向量机分类算法、邻域粗糙集分

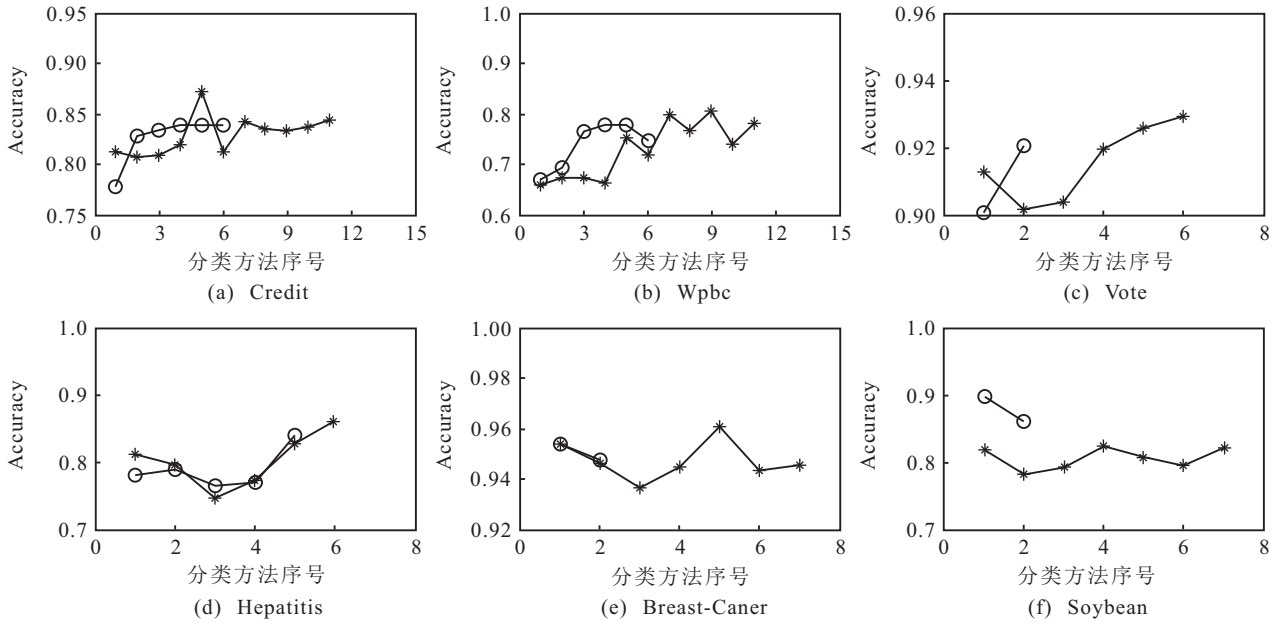

图 2　本文分类结果与相关文献对比图

类算法、K-最近邻分类算法和本文基于覆盖约简的双邻域粗糙集规则学习分类算法(距离度量函数采用联系度距离,取最优精度).

Vote: 图2中星号曲线横轴对应的序号1～5依次表示文献[25]中基于(FSCE、SetCover、PR、IFSPA-IPR、SFFSNTCE)属性约简的朴素贝叶斯分类算法; 6表示本文基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法,本数据集中该分类方法获得的平均精度和最优精度相同,故只取其一. 图2中圆圈曲线横轴对应的序号1、2依次表示朴素贝叶斯分类算法和本文基于覆盖约简的双邻域粗糙集规则学习分类算法(距离度量函数采用联系度距离,取最优精度).

Hepatitis: 图2中星号曲线横轴对应的序号1～4依次表示文献[26]中基于拓展邻域关系属性约简的(决策树、支持向量机、邻域粗糙集、K-最近邻)分类算法; 5、6表示本文基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法,前者表示平均精度,后者表示最优精度. 图2中圆圈曲线横轴对应的序号1～5依次表示决策树分类算法、支持向量机分类算法、邻域粗糙集分类算法、K最近邻分类算法和本文基于覆盖约简的双邻域粗糙集规则学习分类算法(距离度量函数采用联系度距离,取最优精度).

Soybean(large)、Breast-Cancer: 图2中星号曲线横轴对应的序号1～6依次表示文献[24]中基于(Atisa1、Atisa2、Atisa3、Drop3、HMN-EI、ICF)样本选择的K-最近邻($K = 3$)分类算法; 7表示本文基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法,本数据集中该分类方法获得的平均精度和最优精度相同,故只取其一. 图2中圆圈曲线横轴对应的序号1、2依次表示K-最近邻($K = 3$)分类算法和本文基于覆盖约简的双邻域粗糙集规则学习分类算法(距离度量函数采用联系度距离,取最优精度).

最后,通过图2可以看出: 对于Vote和Hepatitis,本文方法的结果最好; 对于Credit和Wpbc,本文方法的结果在10种方法中排名第2、第3,而且对比方法中不乏决策树、最近邻、支持向量机等相关优秀的分类算法; 对于Soybean,本文方法的结果与最优结果基本相当; 对于Breast-Cancer,本文方法的结果与对比文献相差不大,接近95%,而且本文的方法是基于相对较少的属性. 值得说明的是,本文的分类方法均没有对缺失数据作任何处理,也没有采用极端值进行代替,完全保障了数据集的客观真实性,分类效果也是真实数据的反映,这是对比的分类方法所不具备的. 同时,本文分类算法是基于规则学习分类,其泛化能力与支持向量机、最近邻等相关优秀分类算法仍然存在一些差距,但是其分类精度已经与它们非常接近或者部分结果已优于它们,这主要得益于联系度距离函数在不确定对象中的度量能力. 未来,将以联系度距离函数作为距离度量,应用到更多优秀的分类模型中.

5　结　论

本文给出了混合值不完备决策系统的双邻域粗糙集规则学习分类框架. 首先,在距离度量方面,针对不完备数据定义了一种有效的不确定距离度量函数——联系度距离函数; 然后,构建基于联系度距离函数的双邻域粗糙集模型,并给出了该模型的属性约

简算法;最后,以联系度距离函数作为距离度量,给出了基于属性约简、覆盖约简的双邻域粗糙集规则学习分类算法. 在数值实验部分,基于7个UCI数据集,通过对比分析验证了联系度距离函数的有效性和处理缺失数据的优势,以及本文分类方法的客观有效性. 将该距离函数应用于更多优秀的分类模型和聚类模型,以更好地对不完备数据集进行处理及应用,将是下一步的主要研究工作.

参考文献(References)

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

[2] Zhang Y L. Relationships between covering-based rough sets and relation-based rough sets[J]. Information Sciences, 2013, 225(4): 55-71.

[3] Lin T Y. Neighborhood systems — A qualitative theory for fuzzy and rough sets[C]. Advances in Machine Intelligence and Soft Computing. Durham: Duke University, 1997: 132-155.

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

[5] Zhu P F, Hu Q H. Adaptive neighborhood granularity selection and combination based on margin distribution optimization[J]. Information Sciences, 2013, 249(249): 1-12.

[6] 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.

[7] Yao P, Lu Y H. Neighborhood rough set and SVM based hybrid credit scoring classifier[J]. Expert Systems with Applications, 2011, 38(9): 11300-11304.

[8] Chen Y M, Zhang Z J, Zheng J Z, et al. Gene selection for tumor classification using neighborhood rough sets and entropy measures[J]. J of Biomedical Informatics, 2017, 67: 59-68.

[9] Zhu P F, Hu Q H, Han Y H, et al. Combining neighborhood separable subspaces for classification via sparsity regularized optimization[J]. Information Sciences, 2016, 370/371: 270-287.

[10] Lin Y J, Li J J, Lin P R, et al. Feature selection via neighborhood multi-granulation fusion[J]. Knowledge-Based Systems, 2014, 67(3): 162-168.

[11] Liu Y, Xie H, Chen Y H, et al. Neighborhood mutual information and its application on hyperspectral band selection for classification[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 157: 140-151.

[12] Li W W, Huang Z Q, Jia X Y, et al. Neighborhood based decision-theoretic rough set models[J]. Int J of Approximate Reasoning, 2016, 69(C): 1-17.

[13] Chen Y M, Zeng Z Q, Zhu Q X, et al. Three-way decision reduction in neighborhood systems[J]. Applied Soft Computing, 2016, 38: 942-954.

[14] Hu Q H, Yu D, Xie Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866-876.

[15] Kumar S U, Inbarani H H. A novel neighborhood rough set based classification approach for medical diagnosis[J]. Procedia Computer Science, 2015, 47: 351-359.

[16] Du Y, Hu Q H, Zhu P F, et al. Rule learning for classification based on neighborhood covering reduction[J]. Information Sciences, 2011, 181(24): 5457-5467.

[17] Yue X D, Chen Y F, Miao D Q, et al. Tri-partition neighborhood covering reduction for robust classification[J]. Int J of Approximate Reasoning, 2017, 83: 371-384.

[18] 何松华, 康婵娟, 鲁敏, 等. 基于邻域组合测度的属性约简方法[J]. 控制与决策, 2016, 31(7): 1225-1230.
(He S H, Kang C J, Lu M, et al. Attribute Reduction Method based on Neighborhood Combination Measure[J]. Control and Decision, 2016, 31(7): 1225-1230.)

[19] Wilson D R, Martinez T R. Improved heterogeneous distance functions[J]. J of Artificial Intelligence Research, 1997, 6(1): 1-34.

[20] Juhola M, Laurikkala J. On metricity of two heterogeneous measures in the presence of missing values[J]. Artificial Intelligence Review, 2007, 28(2): 163-178.

[21] Gu D X, Liang C Y, Zhao H M. A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis[J]. Artificial Intelligence in Medicine, 2017, 77: 31-47.

[22] Spencer M, Bates Prins S, Beckom M. Heterogeneous distance measures and nearest-neighbor classification in an ecological setting[J]. Missouri J of Mathematical Sciences, 2010, 2(22): 108-123.

[23] Li C Q, Li H W. Correlation weighted heterogeneous euclidean-overlap metric[J]. Int J of Computers and Applications, 2011, 33(4): 341-346.

[24] Cavalcanti G D C, Ren T I, Pereira C L. ATISA: Adaptive threshold-based instance selection algorithm[J]. Expert Systems with Applications, 2013, 40: 6894-6900.

[25] Zhao H, Qin K Y. Mixed feature selection in incomplete decision table[J]. Knowledge-Based Systems, 2014, 57(2): 181-190.

[26] Xu X Y, Liu H F, Shen X F. The research of attribute reduction algorithm based on extension neighborhood relation[J]. J of Computational Information Systems, 2013, 9(16): 6613-6620.

[27] 黄恒秋, 曾玲. 混合值不完备决策信息系统的粗糙分类方法[J]. 计算机工程与应用, 2011, 47(28): 48-51.
(Huang H Q, Zeng L. Rough classification method in incomplete decision information system with hybrid value[J]. Computer Engineering and Applications, 2011, 47(28): 48-51.)

(责任编辑：齐 霁)