ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.3, March 2008, pp.640−649
DOI: 10.3724/SP.J.1001.2008.00640

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于邻域粒化和粗糙逼近的数值属性约简*

胡清华+, 于达仁, 谢宗霞

(哈尔滨工业大学 能源科学与工程学院,黑龙江 哈尔滨 150001)

Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation

HU Qing-Hua+, YU Da-Ren, XIE Zong-Xia

(College of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-451-86413241 ext 252, Fax: +86-451-86413241 ext 221, E-mail: huqinghua@hcms.hit.edu.cn

Hu QH, Yu DR, Xie ZX. Numerical attribute reduction based on neighborhood granulation and rough approximation. *Journal of Software*, 2008,19(3):640−649. http://www.jos.org.cn/1000-9825/19/640.htm

Abstract: To deal with numerical features, a neighborhood rough set model is proposed based on the definitions of δ neighborhood and neighborhood relations in metric spaces. Each object in the universe is assigned with a neighborhood subset, called neighborhood granule. The family of neighborhood granules forms a concept system to approximate an arbitrary subset in the universe with two unions of neighborhood granules: lower approximation and upper approximation. Thereby, the concepts of neighborhood information systems and neighborhood decision tables are introduced. The properties of the model are discussed. Furthermore, the dependency function is used to evaluate the significance of numerical attributes and a forward greedy numerical attribute reduction algorithm is constructed. Experimental results with UCI data sets show that the neighborhood model can select a few attributes but keep, even improve classification power.

Key words: numerical feature; granular computing; neighborhood relation; rough set; variable precision; attribute reduction; feature selection

摘 要: 对于空间中的任一子集,通过基本邻域信息粒子进行逼近,由此提出了邻域信息系统和邻域决策表模型.分析了该模型的性质,并且基于此模型构造了数值型属性的选择算法.利用 UCI 标准数据集与现有算法进行了比较分析,实验结果表明,该模型可以选择较少的特征而保持或改善分类能力.

关键词: 数值特征;粒度计算;邻域关系;粗糙集;可变精度;属性约简;特征选择

中图法分类号: TP18 **文献标识码**: A

Pawlak 教授于 1982 年提出来的粗糙集理论[1]将研究对象的全体称为论域,采用等价关系将论域粒化为若干互斥的等价类,作为描述论域中任意概念的基本信息粒子.对于空间中的任意概念,定义了两个等价类的并

* Supported by the National Natural Science Foundation of China under Grant No.60703013 (国家自然科学基金); the Development Program for Outstanding Young Teachers in Harbin Institute of Technology of China under Grant HITQNJS.2007.017 (哈尔滨工业大学优秀青年教师培养计划); the Scientific Research Foundation of Harbin Institute Technology of China under Grant No.HIT2003.35 (哈尔滨工业大学校基金)

Received 2006-09-18; Accepted 2006-11-27

集:下近似和上近似来逼近这一概念.这一方法自然地模拟了人类的学习和推理过程,学习得到的知识采用产生式规则表示、容易被用户理解、接受和使用,因此得到了广泛的重视,被应用于知识依赖性发现、属性子集选择[2]、决策规则发现[3]、分类分析[4]等问题.

但是,作为一种有效的粒度计算模型,Pawlak 粗糙集定义在经典的等价关系和等价类基础上,只适合处理名义型变量,对于现实应用中广泛存在的数值型数据却不能直接处理.在金融、医疗、科研和工程应用领域,数值型变量无处不在,如电力系统性能分析和设备状态监测与诊断,面临着处理大量数值型数据,包括振动分析中的频谱信号[5],变压器状态分析中的温度、电流、电压信号[6,7],配电系统诊断的电流信号等[8].研究人员在引入粗糙集等机器学习方法来处理该类数据时,往往采用离散化算法把数值型属性转化为符号型属性[9].这一转换不可避免地带来了信息损失[10],计算处理的结果在很大程度上取决于离散化的效果.为了解决这一问题,人们引入了模糊粗糙集模型[11−14]、相似关系粗糙集模型[15]和邻域关系模型.不同的模型基于不同的粒化和逼近机制,都属于粒度计算的研究范畴[16].

粒度计算是模拟人类智能中这一重要能力而提出来的新型计算范式.该方法的基本理念在很多现有的计算方法中普遍应用,例如数值量化、区间分析、机器学习中的分而治之方法等,但正式提出信息粒化和粒度计算的概念要追溯到 Zadeh 教授[16,17]的论文.在这些文献中,给出了为什么要研究粒度计算和如何进行粒度计算的基本框架.由于在信息和知识社会中,人类可获得的信息呈指数形式增长,对信息进行抽象、给出可理解的粒化的信息将有助于用户更好地理解和把握纷繁复杂的信息资源中的有效信息,因此,粒度计算在数据挖掘和知识发现中就显得更为重要.粒度计算的概念一经提出便获得了广泛的认可,吸引了来自人工智能领域、应用数学领域以及工程应用领域的研究人员分析粒度计算的哲学基础、逻辑基础、粒化方法和逼近机制等.词计算[18]、商空间理论[19]、粗糙集理论等都可以理解为一种广义的粒度计算模型.

Lin 于 1988 年提出了邻域模型的概念.该模型可以通过空间中点的邻域来粒化论域,将邻域理解为基本信息粒子,用来描述空间中的其他概念[20].Yao[21]于 1998 年以及 Wu[22]于 2002 年分别研究了 1-step 邻域和 k-step 邻域信息系统的性质.但作者没有进一步分析如何利用所提出的模型进行现实问题的推理.

总体来说,作为一种数值信息粒度计算模型,邻域系统并没有得到足够的重视.本文将系统地分析如何利用拓扑空间中球形邻域的概念,构造基于邻域粗糙集模型的数值数据特征选择算法.该方法直观、易于理解,能够直接处理数值型属性,而无须对其进行离散化处理.因此,与经典粗糙集方法相比,该方法省去了离散化的过程;与 k-step 邻域模型相比,无须建立样本空间中的邻域图.本文采用 UCI 国际标准数据集验证了该方法的有效性.实验发现,适当设置邻域模型的某些参数,可以获得十分理想的效果,系统能够选择很少的特征,却获得与原始数据相当甚至更高的分类精度.与离散化方法[5]和模糊信息熵方法[14]相比,邻域模型方法无论在选择的特征数量和分类精度方面都有较大的优势.该模型可以直接分析数值型数据,从而拓展经典粗糙集理论的应用范围.

1 数值空间的粒化与逼近

粒化和逼近是粗糙集理论和粒度计算的基本问题.Pawlak 粗糙集模型建立在离散空间的分明等价关系之上,等价关系对论域的划分形成了论域空间的粒化.然而对于实数空间而言,对象的取值是连续的.对于此类空间,采用等价关系将导致对个别数值属性的过拟合.邻域结构和序结构是实数空间的重要结构,本文的工作建立在邻域结构的基础之上.

1.1 基于邻域的粒化

邻域有两种定义方法:一种是由邻域内所含对象的数量而定,如经典的 k-近邻方法;另一种是根据在某一度量上邻域中心点到边界的最大距离进行定义.本文采用的是第 2 种方法.

定义 1. 给定一个 N 维的实数空间$\Omega,\Delta:R^N\times R^N\to R$,我们称$\Delta$是 R^N 上的一个度量,如果Δ满足:

(1) $\Delta(x_1,x_2)\geq 0,\Delta(x_1,x_2)=0$,当且仅当 $x_1=x_2,\forall x_1,x_2\in R^N$;

(2) $\Delta(x_1,x_2)=\Delta(x_2,x_1),\forall x_1,x_2\in R^N$;

(3) $\Delta(x_1,x_3)\leq\Delta(x_1,x_2)+\Delta(x_2,x_3),\forall x_1,x_2,x_3\in R^N$,

我们称$\langle \Omega,\Delta\rangle$为度量空间.欧氏距离是实数空间上常用的度量.

定义 2. 给定实数空间上的非空有限集合 $U=\{x_1,x_2,\ldots,x_n\}$,对于 U 上的任意对象 x_i,定义其δ-邻域为

$$\delta(x_i)=\{x|x\in U,\Delta(x,x_i)\le\delta\},$$

其中,$\delta\ge 0$.$\delta(x_i)$称为由 x_i 生成的δ邻域信息粒子,简称为 x_i 的邻域粒子.就二维实数空间而言,基于 1 范数、2 范数和无穷范数的邻域如图 1 所示,分别为菱形、圆形和正方形区域.

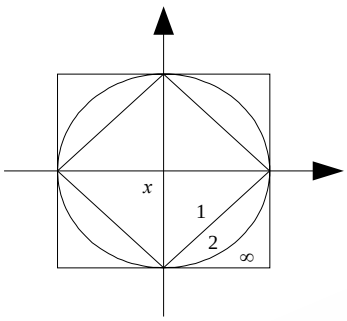

Fig.1 Neighborhood granules in 2-D spaces

图 1 2 维空间邻域粒子

根据度量的性质,我们有:

(1) $\delta(x_i)\ne\varnothing$,因为 $x_i\in\delta(x_i)$;

(2) $x_j\in\delta(x_i)\Rightarrow x_i\in\delta(x_j)$;

(3) $\bigcup_{i=1}^{n}\delta(x_i)=U$.

因此,邻域粒子族$\{\delta(x_i)|i=1,2,\ldots,n\}$构成了 U 的一个覆盖.

邻域信息粒子族引导出论域空间 U 上的一个邻域关系 N,该关系可由一个关系矩阵来表示 $M(N)=(r_{ij})_{n\times n}$,其中,如果 $x_j\in\delta(x_i)$,则 $r_{ij}=1$;否则,$r_{ij}=0$.对于邻域关系 N,我们有:

(1) $r_{ij}=1$;

(2) $r_{ij}=r_{ji}$.

给定一个度量空间$\langle \Omega,\Delta\rangle$,一个非空有限集合 $U=\{x_1,x_2,\ldots,x_n\}$,如果$\delta_1\le\delta_2$,则有:

(1) $\forall x_i\in U:\delta(x_i)\subseteq\delta_2(x_i)$;

(2) $N_1\subseteq N_2$.

论域中所有对象的邻域形成了论域的粒化,邻域粒子族构成了论域空间中的基本概念系统.通过这些基本概念,我们可以逼近空间中的任意概念.

1.2 邻域粗糙集逼近

定义 3. 给定实数空间上的非空有限集合 $U=\{x_1,x_2,\ldots,x_n\}$和 U 上的邻域关系 N,我们称二元组 $NAS=\langle U,N\rangle$为一个邻域近似空间.

定义 4. 给定 $NAS=\langle U,N\rangle$和 $X\subseteq U$,X 在邻域近似空间 $NAS=\langle U,N\rangle$的下近似与上近似分别定义为

$$\underline{N}X=\{x_i\,|\,\delta(x_i)\subseteq X,x_i\in U\},$$

$$\overline{N}X=\{x_i\,|\,\delta(x_i)\cap X\ne\varnothing,x_i\in U\}.$$

$\forall X\in U$, $\overline{N}X\supseteq X\supseteq\underline{N}X$,并定义 X 的近似边界为

$$BN(X)=\overline{N}X-\underline{N}X.$$

$\underline{N}X$ 也称为 X 在近似空间 $NAS=\langle U,N\rangle$中的正域,它是能够被 X 所完全包含的邻域信息粒子的最大并集;相应地, $\overline{N}X$ 是能够完全包含 X 的邻域信息粒子的最小并集.那些与 X 完全无关的邻域信息粒子称为 X 的负域,定义为

$$NEG(X)=U-\overline{N}X=\{x_i\,|\,\delta(x_i)\cap X=\varnothing\}.$$

由于应用中的数据不可避免地会存在一些噪声,以上的定义要求邻域信息粒子要么严格地被 X 所包含,要么严格地与 X 互斥,不存在噪声容忍能力.为此,可以放松对严格包含和严格互斥的要求,而定义程度包含和程度互斥.

定义 5. 设 A,B 是两个集合,定义 A 包含于 B 的程度 $I(A,B)$为

$$I(A,B)=\frac{Card(A\cap B)}{Card(A)},\ A,B\ne\varnothing.$$

当 $A=\varnothing$时,规定 $I(A,B)=0$.我们有 $0\le I(A,B)\le 1$.

定义 6. 定义可变精度 k-下近似和上近似分别为

$$N^k X=\{x_i \mid I(\delta(x_i),X)\geq k,x_i\in U\},$$

$$\overline{N^k}X=\{x_i \mid I(\delta(x_i),X)\geq 1-k,x_i\in U\}.$$

其中,$k\geq 0.5$.

1.3 邻域决策系统

对于很多模式识别问题,给定的样本由一系列数值型特征描述,样本被分成少数的几个类别.分类学习就是从这样的样本集合中学习一个函数,实现从特征空间到决策的映射.此类问题可以表示为如下形式:

定义 7. 给定样本集合 $U=\{x_1,x_2,\ldots,x_n\}$,A 是描述 U 的实数型特征集合,D 是决策属性.如果 A 生成论域上的一族邻域关系,则称 $NDT=\langle U,A,D\rangle$为一个邻域决策系统.

举例来讲,$U=\{x_1,x_2,x_3,x_4,x_5\}$,$a$ 是 U 的一个属性,$f(x,a)$表示样本 x 在属性 a 上的取值.$f(x_1,a)=1.1$,$f(x_2,a)=1.2$,$f(x_3,a)=1.6$,$f(x_4,a)=1.8$,$f(x_5,a)=1.9$.当指定邻域大小为 0.2 时,由于$|f(x_1,a)-f(x_2,a)|\leq 0.2$,则 $x_2\in\delta(x_1)$,$x_1\in\delta(x_2)$.类似地可计算$\delta(x_1)=\{x_1,x_2\}$,$\delta(x_2)=\{x_1,x_2\}$,$\delta(x_3)=\{x_3,x_4\}$,$\delta(x_4)=\{x_3,x_4,x_5\}$,$\delta(x_5)=\{x_4,x_5\}$.如果存在多个属性,我们以此可计算样本之间的距离,并进而计算样本的邻域.

定义 8. 给定一个邻域决策系统 $NDT=\langle U,A,D\rangle$,D 将 U 划分为 N 个等价类:$X_1,X_2,\ldots,X_N$,$\forall B\subseteq A$,定义决策 D 关于 B 的下近似和上近似为

$$\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i,$$

$$\overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i,$$

其中,

$$\underline{N_B}X=\{x_i \mid \delta_B(x_i)\subseteq X,x_i\in U\},$$

$$\overline{N_B}X=\{x_i \mid \delta_B(x_i)\cap X\neq\varnothing,x_i\in U\}.$$

$\delta_B(x_i)$是由属性 B 和度量Δ生成的邻域信息粒子.

决策的边界定义为

$$BN(D)=\overline{N_B}D-\underline{N_B}D.$$

图 2 给出了一个二维空间的两类分类问题,以“*”标识的样本为第 1 类样本,以“+”标识的为第 2 类样本.我们看出,样本 x_1 的圆形邻域内的样本都来自于第 1 类,因此,x_1 属于第 1 类的下近似;x_3 的邻域内样本都来自第 2 类,因此,x_3 属于第 2 类的下近似;样本 x_2 的邻域内既有第 1 类的样本,也有第 2 类的样本,因此,x_2 是边界样本.我们可以看出,这一定义方式与我们对分类问题的直观认识是一致的.图 3 给出了经典粗糙集模型的几何解释,深色标识的等价类完全属于 X,它们属于 X 的下近似.与邻域粗糙集模型对照,这两种模型是一致的.

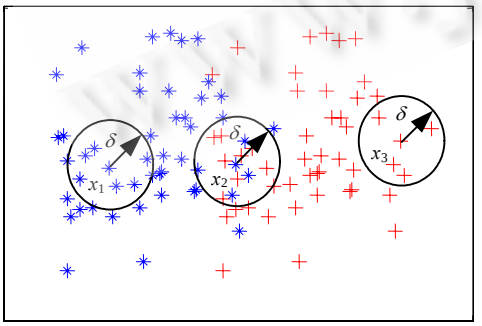

Fig.2 Neighborhood rough sets in real spaces

图 2 实数空间的邻域粗糙集

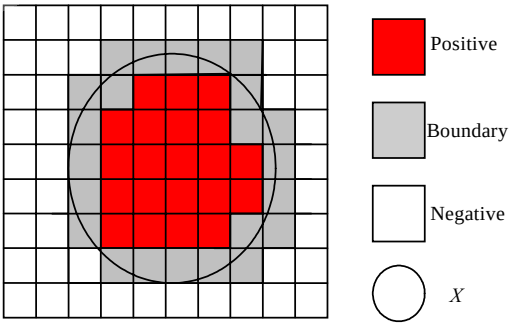

Fig.3 Classical rough sets in discrete spaces

图 3 离散空间的经典粗糙集

决策 D 的下近似也称为决策正域,记为 $POS_B(D)$.正域的大小反映了分类问题在给定属性空间中的可分离程度.正域越大,表明各类的重叠区域即边界越少.我们可以根据此属性集,更加精细地描述此分类问题.因此,定义决策属性 D 对条件属性 B 的依赖性为

$$\gamma_B(D)=Card(\underline{N_B}D)/Card(U).$$

记为 $B\Rightarrow_\gamma D$.$0\le\gamma_B(D)\le 1$,表示了样本集合中根据条件属性 B 的描述,那些能够被某一类决策完全包含的样本所占全体样本的比率.显然,正域越大,决策 D 对条件 B 的依赖性越强.

定理 1. 给定邻域信息系统$\langle U,A,D\rangle$,$B\subseteq A$,有

$$POS_B(D)\cap BN_B(D)=\varnothing;POS_B(D)\cup BN_B(D)=U.$$

证明:假设决策属性将样本分为 $D_1,D_2,\ldots,D_N$.如果 $k\ne l$,我们有 $D_k\cap D_l=\varnothing$.不失一般性,分析任意一个样本 $x_i\in U$,如果$\delta(x_i)\subseteq D_k$,则$\delta(x_i)\cap D_l=\varnothing(l\ne k)$,从而 $x_i\in\underline{N_B}D_l$, $x_i\notin\overline{N_B}D_l$.由此可知,对任意一个样本,如果它属于某一决策的正域,则它一定不会属于任一决策的边界.从而可知,$POS_B(D)\cap BN_B(D)=\varnothing$.由于 $\bigcup_{i=1}^{n}D_i=U$,任意一个 $x_i\in U$,必然存在 D_j,使得$\delta(x_i)\cap D_j\ne\varnothing$,因此, $x_i\in\overline{N_B}D_j$,则 $x_i\in\underline{N_B}D_j$ 或者 $x_i\in BN_B(D_j)$. □

定理 2. 给定度量Δ和邻域大小阈值δ,如果 $B_1\subseteq B_2$,且 $x_i\in POS_{B_1}(D)$,则 $x_i\in POS_{B_2}(D)$ 成立.

证明:不失一般性,我们假设 $x_i\in\underline{B_1}D_j$,其中 D_j 表示决策类别为 j 的样本集,即 $\delta_{B_1}(x_i)\subseteq D_j$.基于同样的邻域阈值$\delta$和度量$\Delta$,我们有 $\delta_{B_2}(x_i)\subseteq\delta_{B_1}(x_i)$,因此, $\delta_{B_2}(x_i)\subseteq\delta_{B_1}(x_i)\subseteq D_j$.从而有 $x_i\in POS_{B_2}(D)$. □

定理 3. $\gamma_B(D)$是单调的,如果 $B_1\subseteq B_2\subseteq\ldots\subseteq C$,则 $\gamma_{B_1}(D)\le\gamma_{B_2}(D)\le\ldots\le\gamma_C(D)$.

证明:基于定理 2 可知,$\forall x_i\in POS_{B_1}(D)$,则我们有 $x_i\in POS_{B_2}(D),\ldots,x_i\in POS_C(D)$.同时,可能存在 $x_j\in U$,使得 $x_j\in BN_{B_1}(D)$,但 $x_j\in POS_{B_2}(D),\ldots,x_j\in POS_C(D)$.因此可得:

$$POS_{B_1}(D)\le POS_{B_2}(D)\le\ldots\le POS_C(D).$$

由于$\gamma_B(D)=Card(POS_B(D))/Card(U)$,我们有 $\gamma_{B_1}(D)\le\gamma_{B_2}(D)\le\ldots\le\gamma_C(D)$. □

定义 9. 给定邻域信息系统$\langle U,A,D\rangle$,$B\subseteq A$,$\forall a\in B$,如果$\gamma_{B-a}(D)<\gamma_B(D)$,我们称 a 相对于 B 而言是必不可少的,否则,$\gamma_{B-a}(D)=\gamma_B(D)$,我们称 a 是多余的.如果$\forall a\in B$ 都是必不可少的,我们称 B 是独立的.

如果$\gamma_{B-a}(D)=\gamma_B(D)$,则表明从系统中去掉属性 a,系统的正域没有发生改变,因此,各类的可区分性不变.此时,属性 a 没有给分类带来任何贡献,因此说 a 是多余的.相反地,如果删除 a,系统的决策正域变小了,则表明各类的可区分性变差了.此时,a 不能被删除.

定义 10. 给定邻域决策系统$\langle U,A,D\rangle$,我们称 $B\subseteq A$ 是 A 的一个约简,如果 B 满足:

(1) $\forall a\in B,\gamma_{B-a}(D)<\gamma_B(D)$;

(2) $\gamma_B(D)=\gamma_A(D)$.

该定义的条件(1)要求在一个约简中不存在多余的属性,所有的属性都应该是必不可少的;条件(2)要求约简不能降低系统的区分能力,约简应该与系统中全部条件属性具有相同的分辨能力.这一定义与经典粗糙集模型中的定义在形式上是完全一致的,然而,由于该模型定义了数值空间中的粒化和逼近,而经典粗糙集是定义在离散空间的,因此适合于完全不同的应用场合.

定义 11. 给定邻域决策系统$\langle U,A,D\rangle$,$B_1,\ldots,B_i,\ldots,B_k$ 是此系统的全部约简,则定义 $Core=\bigcap_{i=1}^{k}B_i$ 为决策系统的核.

2 基于邻域模型的前向贪心数值属性约简

发现决策系统的全部约简,需要测试 2^N-1 个属性子集,以检验它们是否满足约简的条件,其中,N 是条件属性的数量.这对于那些为数十个、甚至上百个属性的决策系统而言,计算量是不可容忍的.本文将基于依赖性函

数,构造一种前向贪心约简算法.

依赖性函数定义了条件属性对分类的贡献,因此,可以作为属性集合重要性的评价指标.

定义 12. 给定一个邻域决策系统 $NDT=\langle U,A,D\rangle$,$B\subseteq A$,$\forall a\in A-B$,定义 a 相对于 B 的重要度为

$$SIG(a,B,D)=\gamma_{B\cup a}(D)-\gamma_B(D).$$

属性的重要性是属性本身、属性相对的属性子集以及决策变量三者构成的一个函数.

基于属性重要度指标,我们可以构造贪心式属性约简算法.该算法以空集为起点,每次计算全部剩余属性的属性重要度,从中选择属性重要度值最大的属性加入约简集合中,直到所有剩余属性的重要度为 0,即加入任何新的属性,系统的依赖性函数值不再发生变化为止.前向搜索算法能够确保重要的属性首先被加入到约简中,从而不损失重要的特征.后向搜索算法却难以保证这个结果,因为对于有大量冗余特征的决策系统而言,即使那些重要的属性被删除也不一定会降低整个系统的区分能力,因此,系统最终可能保留了大量区分能力很弱、但作为一个整体依然能够保持原始数据的分辨能力的特征,而不是少量区分能力很强的特征.基于邻域粗糙集模型的数值属性约简算法描述见算法 1.

算法 1. 基于邻域粗糙集模型的数值特征选择.

Input: $NDT=\langle U,A,D\rangle$;

Output: 约简 red.

Step 1: $\forall a\in A$:计算邻域关系 N_a;

Step 2: $\varnothing\rightarrow red$;

Step 3: 对任意 $a_i\in A-red$

计算 $SIG(a_i,red,D)=\gamma_{red\cup a}(D)-\gamma_{red}(D)$, //此处定义$\gamma_{\varnothing}(D)=0$

Step 4: 选择 a_k,其满足:

$$SIG(a_k,red,D)=\max_i(SIG(a_i,red,D))$$

Step 5: If $SIG(a_k,red,D)>0$,

$red\cup a_k\rightarrow red$

go to Step 3

else

return red, end

3 实验分析

为了验证该方法的有效性,我们从 UCI 数据集中挑选了 6 组数据,描述见表 1.可以发现,这 6 个分类问题的条件属性全部是数值型的.

Table 1 Data description

表 1 数据描述

	Data set	Abbreviation	Samples	Numerical features	Classes
1	Ionosphere	Iono	351	34	2
2	Sonar, Mines vs. Rocks	Sonar	208	60	2
3	Small soybean	Soy	47	35	4
4	Wisconsin diagnostic breast cancer	WDBC	569	31	2
5	Wisconsin prognostic breast cancer	WPBC	198	33	2
6	Wine recognition	Wine	178	13	3

我们将比较本文的方法、基于 FCM 离散化方法与模糊信息熵方法[14]在选择特征数量和分类精度之间的差别.为了比较所选择特征的分类能力,我们引入流行的 CART 和 RBF-SVM 分类学习算法,以 10 折交叉验证的分类精度来评价选择特征的质量.其中,$N1$ 和 $N2$ 分别表示原始特征数量,Accuracy1 和 Accuracy2 分别表示原始特征的分类精度和选择的特征子集的分类精度.

在计算各样本的邻域时,所有数值型属性被标准化到[0,1]区间,以减少因各属性量纲不一致对结果的影响.同时,我们设置δ=0.125,即邻域的直径为 0.25.随后,我们会展示邻域大小对属性约简结果的影响.

表 2 展示了采用离散化和经典粗糙集属性约简相结合得到的特征数量、分类精度与原始数据的比较.表 3 给出的是采用模糊信息熵直接约简数值属性的特征数和分类精度.表 4 和表 5 是基于邻域粗糙集模型和可变精

度邻域粗糙集模型的计算结果,其中,邻域的直径为 0.25.在可变精度的计算中,我们改变可变精度的阈值,然后记录最好分类精度时对应的特征数.

Table 2 Number and accuracy of selected features based on FCM and classical rough sets (dividing into 4 intervals for each attribute)

表 2 基于 FCM 离散化和粗糙集结合的特征选择方法选择的特征数量和分类精度(离散为 4 个区间)

	Feature		CART		RBF-SVM	
Data	*N*1	*N*2	Accuracy 1	Accuracy 2	Accuracy 1	Accuracy 2
Iono	34	10	0.8755±0.0693	0.9089±0.0481	0.9379±0.0507	0.9348±0.0479
Sonar	60	6	0.7207±0.1394	0.6926±0.0863	0.8510±0.0948	0.7074±0.1004
Soy	35	2	0.9750±0.0791	1.0000±0.0000	0.9300±0.1135	1.0000±0.0000
Wdbc	31	8	0.9050±0.0455	0.9351±0.0339	0.9808±0.0225	0.9649±0.0183
Wpbc	33	7	0.6963±0.0826	0.6955±0.1018	0.7779±0.0420	0.7837±0.0506
Wine	13	4	0.8986±0.0635	0.8972±0.0741	0.9889±0.0234	0.9486±0.0507
Average	34.33	6.17	0.8452	0.8549	0.9111	0.8899

Table 3 Feature selection based on fuzzy entropy

表 3 基于模糊信息熵的特征选择

	Feature		CART		RBF-SVM	
Data	*N*1	*N*2	Accuracy 1	Accuracy 2	Accuracy 1	Accuracy 2
Iono	34	13	0.8755±0.0693	0.9068±0.0564	0.9379±0.0507	0.9462±0.0365
Sonar	60	12	0.7207±0.1394	0.7160±0.0857	0.8510±0.0948	0.8271±0.0902
Soy	35	2	0.9750±0.0791	1.0000±0.0000	0.9300±0.1135	1.0000±0.0000
Wdbc	31	17	0.9050±0.0455	0.9193±0.0318	0.9808±0.0225	0.9702±0.0248
Wpbc	33	17	0.6963±0.0826	0.7103±0.1092	0.7779±0.0420	0.8087±0.0601
Wine	13	9	0.8986±0.0635	0.9097±0.0605	0.9889±0.0234	0.9833±0.0268
Average	34.33	11.67	0.8452	0.8603	0.9111	0.9226

Table 4 Feature selection based on neighborhood model (attribute values in [0, 1] and neighborhood size is 0.25)

表 4 基于邻域粗糙集模型的特征选择(变量标准化到[0,1]区间,邻域宽度为 0.25)

	Feature		CART		RBF-SVM	
Data	*N*1	*N*2	Accuracy 1	Accuracy 2	Accuracy 1	Accuracy 2
Iono	34	12	0.8755±0.0693	0.9063±0.0396	0.9379±0.0507	0.9293±0.0627
Sonar	60	7	0.7207±0.1394	0.7550±0.0683	0.8510±0.0948	0.8364±0.0837
Soy	35	2	0.9750±0.0791	1.0000±0.0000	0.9300±0.1135	1.0000±0.0000
Wdbc	31	21	0.9050±0.0455	0.9228±0.0361	0.9808±0.0225	0.9790±0.0161
Wpbc	33	11	0.6963±0.0826	0.6453±0.1292	0.7779±0.0420	0.7842±0.0769
Wine	13	6	0.8986±0.0635	0.9208±0.0481	0.9889±0.0234	0.9833±0.0268
Average	34.33	9.83	0.8452	0.8584	0.9111	0.9187

Table 5 Feature number with the best classification accuracies in variable precision neighborhood rough sets

表 5 可变精度邻域下最好的分类精度对应的特征数

	CART		SVM	
Data	Features	Accuracies	Features	Accuracies
Iono	7	0.9236±0.0613	8	0.9400±0.0373
Sonar	7	0.7876±0.0752	7	0.8364±0.0837
Soy	2	1.0000±0.0000	2	1.0000±0.0000
Wdbc	7	0.9369±0.0340	21	0.9790±0.0160
Wpbc	1	0.7484±0.0862	10	0.8042±0.0711
wine	6	0.9438±0.0371	6	0.9833±0.0268
Average	5	0.8900	9	0.9238

我们可以发现,这几种算法都能有效地降低特征数量.相对而言,离散化算法得到的特征数量较少,但分类精度也较低,模糊熵方法虽然得到较高的分类精度,但保留了较多的特征.可变精度邻域粗糙集不仅获得了最高的分类精度,而且保留的特征也是最少的.

图 4~图 6 分别展示了约简中特征数量、分类精度随邻域的大小δ和可变精度阈值 k 的变化情况,其中,δ的

取值以 0.05 为步长从 0 到 1 变化,k 的取值以 0.05 为步长从 0.5 到 1 变化.从图中可以发现,当δ<0.1 时,约简算法发现的特征很少,相应的分类精度也较低;当δ>0.8 时,系统发现的特征也很少.另外一个区域是当δ和 k 的取值都较大时,即图中的右上角区域,算法选中的特征少,相应的分类精度低,因此,这些区域的参数是不宜采用的.相对而言,参数δ在[0.2,0.4]之间取值,k 在[0.8,0.9]之间较为理想.

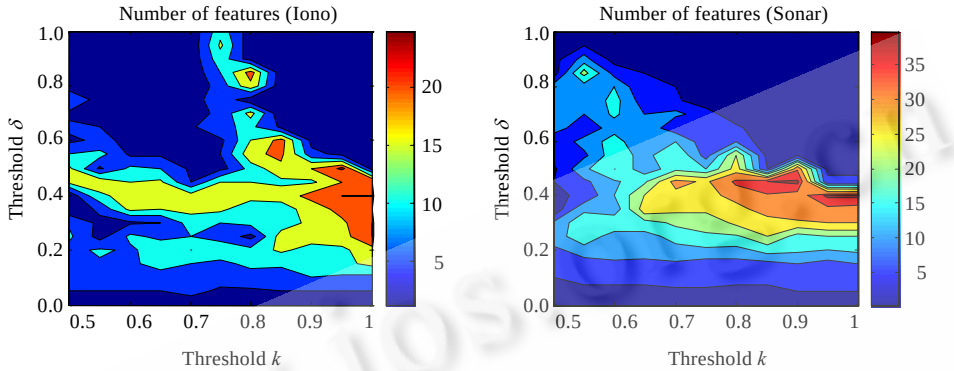

Fig.4 Number of selected features varying with the size of neighborhood δ and the threshold k

图 4 约简中特征数量随邻域大小δ和可变精度粗糙集阈值 k 的变化

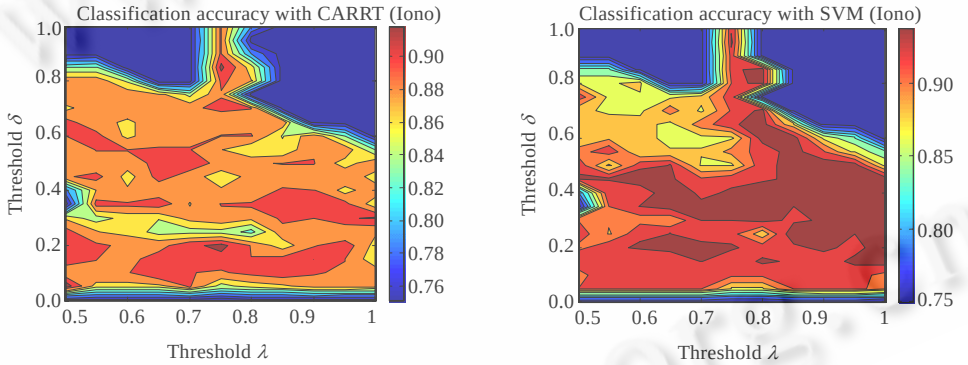

Fig.5 Accuracy of selected features varying with the size of neighborhood δ and the threshold k (Iono data)

图 5 约简分类能力随邻域大小δ和可变精度粗糙集阈值 k 的变化(Iono)

4 结论与展望

本文基于度量空间邻域和邻域关系的概念提出了一种实数空间的粗糙集模型.该模型以论域空间中任意对象的邻域形成论域空间的粒化,以下近似和上近似来逼近空间中的任一概念,从而实现了实数空间的粒度计算.进一步地,我们展示了该模型在分类问题数值型属性约简选择中的应用.实验分析表明,邻域粗糙集模型可以选择少量的特征,并且保持甚至显著提高约简数据的分类精度,验证了该方法的有效性.

邻域粗糙集模型采用了范数计算样本之间的相似性,当我们将邻域的大小设为 0 时,则要求邻域内的样本在所有特征上的取值都相等,此时,邻域粒子变成了等价信息粒子,邻域模型退化为经典粗糙集模型.因此,当数据中存在符号型和数值型变量时,我们只需设置符号型变量的邻域大小为 0,数值型变量为大于 0 的常数,则该模型即可处理混合变量数据集.后续的工作将沿着这个方向开展,研究该模型用于混合变量系统的性能和参数设置方法.此外,基于邻域粗糙集模型的海量数据快速约简算法也是研究的方向之一[24].

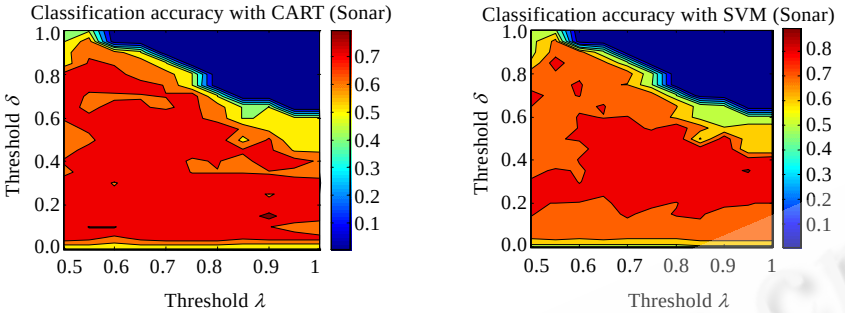

Fig.6 Accuracy of selected features varying with the size of neighborhood δ and the threshold k (Sonar data)

图 6 约简分类能力随邻域大小 delta 和可变精度粗糙集阈值 k 的变化(Sonar)

References:

[1] Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic, 1991.

[2] Wang J, Wang R, Miao DQ, *et al*. Data enriching based on rough set theory. Chinese Journal of Computers, 1998,21(5):393−400 (in Chinese with English abstract).

[3] Chang LY, Wang GY, Wu Y. An approach for attribute reduction and rule generation based on rough set theory. Journal of Software, 1999,10(11):1207−1211 (in Chinese with English abstract).

[4] Shi Y, Sun YF, Zuo C. Spatial data classification based on rough set. Journal of Software, 2000,11(5):673−678 (in Chinese with English abstract).

[5] Yu DR, Hu QH, Bao W. Combining rough set methodology and fuzzy clustering for knowledge discovery from quantitative data. Proc. of the Chinese Society for Electrical Engineering, 2004,24(6):205−210 (in Chinese with English abstract).

[6] Zhu YL, Wu LZ, Li XY. Synthesized diagnosis on transformer faults based on Bayesian classifier and rough set. Proc. of the Chinese Society for Electrical Engineering, 2005,25(10):159−165 (in Chinese with English abstract).

[7] Wang YQ, Lü FC, Li HM. Synthetic fault diagnosis method of power transformer based on rough set theory and Bayesian network. Proc. of the Chinese Society for Electrical Engineering, 2006,26(8):137−141 (in Chinese with English abstract).

[8] Sun QY, Zhang HG. Fault diagnose algorithm of distribution system by continuous signals based on rough sets. Proc. of the Chinese Society for Electrical Engineering, 2006,26(11):156−161 (in Chinese with English abstract).

[9] Xie H, Cheng HZ, Niu DX. Discretization of continuous attributes in rough set theory based on information entropy. Chinese Journal of Computers, 2005,28(9):1570−1574 (in Chinese with English abstract).

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

[11] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. Int'l Journal General Systems, 1990,17:191−209.

[12] Hu QH, Yu DR, Xie ZX, Liu JF. Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans. on Fuzzy Systems, 2006,14(2):191−201.

[13] Yeung DS, Chen DG, Tsang ECC, Lee JWT, Wang XZ. On the generalization of fuzzy rough sets. IEEE Trans. on Fuzzy Systems, 2005,13(3):343−361.

[14] Hu QH, Yu DR, Xie ZX. Information-Preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognition Letters, 2006,27(5):414−423.

[15] Slowinski R, Vanderpooten D. A generalized definition of rough approximations based on similarity. IEEE Trans. on Knowledge and Data Engineering, 2000,12(2):331−336.

[16] Zadeh LA. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 1997,19:111−127.

[17] Zadeh LA. Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yager RR, eds. Advances in Fuzzy Set Theory and Applications. Amsterdam, 1979. 3−18.

[18] Zadeh LA. Fuzzy logic=Computing with words. IEEE Trans. on Fuzzy Systems, 1996,4(2):103−111.

[19] Zhang L, Zhang B. Theory of fuzzy quotient space (methods of fuzzy granular computing). Journal of Software, 2003,14(4): 770−776 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/770.htm

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

[21] Yao YY. Relational interpretation of neighborhood operators and rough set approximation operators. Information Sciences, 1998, 111(198):239−259.

[22] Wu WZ, Zhang WX. Neighborhood operator systems and approximations. Information Sciences, 2002,144(1-4):201−217.

[23] Liu Q, Liu SH, Zheng F. Rough logic and its application in data reduction. Journal of Software, 2001,12(3):415−419 (in Chinese with English abstract).

[24] Xu ZY, Liu ZP, Yang BR, Song W. A quick attribute reduction algorithm with complexity of max($O(|C||U|)$,$O(|C|^2|U/C|)$). Chinese Journal of Computers, 2006,29(3):391−399 (in Chinese with English abstract).

附中文参考文献:

[2] 王珏,王任,苗夺谦等.基于 Rough Set 理论的“数据浓缩”.计算机学报,1998,21(5):393−400.

[3] 常犁云,王国胤,吴渝.一种基于 Rough Set 理论的属性约简及规则提取方法.软件学报,1999,10(11):1207−1211.

[4] 石云,孙玉芳,左春.基于 Rough Set 的空间数据分类方法.软件学报,2000,11(5):673−678.

[5] 于达仁,胡清华,鲍文.融合粗糙集和模糊聚类的连续数据知识发现.中国电机工程学报,2004,24(6):205−210.

[6] 朱永利,吴立增,李雪玉.贝叶斯分类器与粗糙集相结合的变压器综合故障诊断.中国电机工程学报,2005,25(10):159−165.

[7] 王永强,律方成,李和明.基于粗糙集理论和贝叶斯网络的电力变压器故障诊断方法.中国电机工程学报,2006,26(8):137−141.

[8] 孙秋野,张化光.基于粗糙集的配电系统连续信号故障诊断方法.中国电机工程学报,2006,26(11):156−161.

[9] 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.计算机学报,2005,28(9):1570−1574.

[19] 张铃,张钹.模糊商空间理论(模糊粒度计算方法).软件学报,2003,14(4):770−776. http://www.jos.org.cn/1000-9825/14/770.htm

[23] 刘清,刘少辉,郑非.Rough 逻辑及其在数据约简中的应用.软件学报,2001,12(3):415−419.

[24] 徐章艳,刘作鹏,杨炳儒,宋威.一个复杂度为 max($O(|C||U|)$,$O(|C|^2|U/C|)$)的快速属性约简算法.计算机学报,2006,29(3):391−399.

胡清华(1976－),男,湖南娄底人,博士生,讲师,主要研究领域为机器学习,粗糙集理论.

于达仁(1966－),男,博士,教授,主要研究领域为智能控制,故障诊断.

谢宗霞(1981－),女,博士生,主要研究领域为机器学习,图像处理.