基于邻域近似条件熵的启发式属性约简*

张 宁, 范年柏
(湖南大学 信息科学与工程学院, 长沙 410082)

摘 要: 目前粗糙集的研究局限于有限集,且现有的邻域粗糙集属性约简算法中属性重要性度量方式单一。针对邻域粗糙集存在的问题,提出了基于无限集的邻域近似条件熵模型。该模型以邻域近似条件熵下的属性重要度为启发条件,构造了一种基于邻域近似条件熵的前向贪心搜索属性约简算法。利用熵的单调性,证明了算法的正确性,并分析了算法的时间复杂度。通过实例分析和多个 UCI 数据集上的实验表明,所提出的算法是可行的,能有效减少属性数量,与现有的算法相比,不仅能够获得较小的属性约简结果,而且具有较好的分类性能。

关键词: 邻域粗糙集; 条件熵; 属性约简; 属性重要性
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2018)05-1395-04
doi:10.3969/j.issn.1001-3695.2018.05.024

Heuristic attribute reduction based on neighborhood approximate conditional entropy

Zhang Ning, Fan Nianbai
(*College of Computer Science & Electronic Engineering*, *Hunan University*, *Changsha* 410082, *China*)

Abstract: The research of the rough set has been limited to the finite set so far, and the attribute significance measurement is single in attribute reduction algorithm based on the neighborhood rough set model. In order to solve the problem in neighborhood rough set, this paper proposed the neighborhood approximate conditional entropy model based on infinite set. Furthermore, it constructed a forward greedy attribute reduction algorithm which used the neighborhood approximate conditional entropy as heuristic condition in this model. The monotonicity of the entropy proved the correctness of the algorithm, and the time complexity was analyzed. The instance analysis and the experimental results of UCI data sets show that the proposed attribute reduction algorithm is feasible and it can reduce the number of attributes effectively. It can get smaller attribute reduction results, and better classification power.

Key words: neighborhood rough set; conditional entropy; attribute reduction; attribute significance

0 引言

粗糙集理论是 Pawlak 教授[1]于 1982 年提出的一种处理模糊和不确定性知识的有效数学工具。属性约简是粗糙集理论的重要研究内容之一。目前已有的属性约简模型主要有基于差别矩阵的属性约简模型[2,3]、代数观下的属性约简模型[4]以及信息观下的属性约简模型[5]。但是,Pawlak 粗糙集定义在等价关系基础上,只适合处理离散型数据,不能直接处理现实生活中广泛存在的数值型数据,经典粗糙集在处理包含数值型数据的决策表的属性约简问题时,常采用离散化方法[6]把数值型数据转换为离散型数据,但离散化过程必定会造成某种程度的信息损失。邻域粗糙集模型的出现为解决这一问题提供了新的思路。利用邻域粗糙集方法对包含数值型数据的决策表属性进行选取,无须对其进行离散化处理,能最大限度地保持数据集的分类能力。

Lin[7]提出了邻域系统模型的概念,在这个模型中使用邻域粒化论域。此后,Wu 等人[8]对邻域信息系统的数学性质进行了研究。胡清华等人[9,10]从代数定义的角度设计了基于邻域粗糙集模型的属性约简及其快速属性约简算法,算法以正区域作为启发式信息。但基于正区域的邻域粗糙集属性约简算法仅考虑被正确区分的样本数,梁海龙等人[11]针对这一问题提出了区分对象集的概念,利用新的属性重要度度量方式构造了一种新的启发式属性约简算法,算法本质上也是基于代数观点的。Hu 等人[12]将 Shannon 信息熵扩展为邻域信息熵,提出了邻域互信息的概念,并以邻域互信息作为属性重要度度量方式构造了信息论观点下的邻域粗糙集属性约简算法。续欣莹等人[13]提出了信息观下基于不一致邻域矩阵的属性约简算法。

但是目前邻域粗糙集讨论的是有限集,具有一定的局限性;并且现有的邻域粗糙集属性约简方法中,通常仅从代数观点或者信息论观点单方面出发。基于代数观点的属性重要度定义仅刻画属性对论域中包含的确定分类子集的影响;而基于信息论观点的属性重要度定义仅刻画属性对论域中包含的不确定分类子集的影响[14,15]。因此单独的两种标准都不是完备的,谢玲玲等人[16]结合邻域粗糙集的代数观点和信息论观点,提出了一种更加全面的条件熵模型和相应算法,但是该模型同样定义在有限集合下,并且文献[16]的算法无法处理无核的系统。

针对上述问题,本文引入测度[17],将邻域粗糙集推广到无穷集;然后结合邻域近似精度和邻域条件熵构造了基于无限集的邻域近似条件熵模型,证明了其单调性。最后以属性重要度为启发函数构造出一种新的启发式属性约简算法。

收稿日期: 2016-12-25; **修回日期**: 2017-03-13 **基金项目**: 湖南省科技计划应用基础研究重点项目(2016JC2014)

作者简介:张宁(1990-),女,山东枣庄人,硕士,主要研究方向为数据挖掘(768138016@qq.com);范年柏(1962-),男,湖南岳阳人,副教授,博士,主要研究方向为形式化方法、数据挖掘.

1 相关概念

用于分类学习的结构化数据可形式化为一个四元组 IS = $\langle U,A,V,f\rangle$。其中,U 为对象的非空无限集合,称为论域;$A=\{a_1,a_2,\cdots,a_m\}$ 为描述对象的全部属性所组成的非空有限集合;$V=\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;$f:U\times A\to V$ 是一个信息函数,它为每个对象的每个属性赋予一个信息值,即 $\forall a\in A, x\in U, f(x,a)\in V_a$。如果属性集 $A=C\cup D$,$C\cap D=\varnothing$,其中 C 为条件属性集,D 为决策属性集,则 $\langle U,A,V,f\rangle$ 是一个决策表。

定义 1[10] 对于任意的 $x_i\in U, B\subseteq C, x_i$ 在属性子集 B 上的 δ-邻域定义为

$$\delta_B(x_i)=\{x_j\mid x_j\in U,\Delta_B(x_i,x_j)\leqslant\delta\}$$

其中 $\delta\geqslant 0$。

常用的距离函数有三种,在 N 维空间 $A=\{a_1,a_2,\cdots,a_n\}$ 中,考虑两个对象 x_1 和 x_2,$f(x,a_i)$ 表示对象 x 在第 a_i 个属性上的值,Minkowsky 距离函数定义为

$$\Delta_P(x_1,x_2)=\left(\sum_{i=1}^{N}|f(x_1,a_i)-f(x_2,a_i)|^P\right)^{1/P}$$

当 $P=1$ 时,此函数称为 Manhattan 距离;当 $P=2$ 时,称为 Euclidean 距离;当 $P=\infty$ 时,称为 Chebychev 距离。

定义 2 给定一个邻域决策系统 NDT = $\langle U,C\cup D,V,f\rangle$,$\{Y_1,Y_2,\cdots,Y_m,\cdots\}$ 是 U 在决策属性 D 上的划分,$B\subseteq C$ 生成 U 上的邻域关系 N_B,$\delta_B(x_i)$ 表示对象 x_i 在属性 B 下的邻域,决策属性集 D 关于 B 的邻域下近似和邻域上近似分别定义为

$$\overline{N_B}D=\cup_{i=1}^{\infty}\overline{N_B}Y_i,\ \underline{N_B}D=\cup_{i=1}^{\infty}\underline{N_B}Y_i$$

其中:$\underline{N_B}Y=\{x_i\mid\delta_B(x_i)\subseteq Y,x_i\in U\}$,$\overline{N_B}Y=\{x_i\mid\delta_B(x_i)\cap Y\neq\varnothing,x_i\in U\}$。

定义 3[17] 设 E 为 R^n 中任一点集,对于每一列覆盖 E 的开球 I_i,给出它的体积总和 $\mu=\sum_i|I_i|$,所有这一切的 μ 组成一个下方有界的数集,它的下确界(由 E 完全确定)称为 E 的勒贝格外测度,简称 L 外测度或外测度,即

$$m^*(E)=\inf_{E\subset\cup I_i}\sum_i|I_i|$$

E 的内测度 $m_*(E)=|I|-m^*(I-E)$。

如果 $m^*E=m_*E$,则称 E 可测,并记做 $m(E)$。当 U 测度为 0 时,则改为势(基数)。为方便起见,统一采用此记号:$m(X)$。

性质 1[17] 如果集合 $X_1\subseteq X_2$,则 $m(X_1)\leqslant m(X_2)$。

定义 4 给定一个邻域决策系统,NDT = $\langle U,C\cup D,V,f\rangle$,对于 $\forall B\subseteq C, X\subseteq U$,$X$ 在邻域关系 N_B 下的邻域近似精度定义为

$$\alpha_B(X)=\frac{m(\underline{N_BX})}{m(\overline{N_BX})}$$

其中 $X\neq\varnothing$。显然,$0\leqslant\alpha_B(X)\leqslant 1$。本文将文献[18]中容差关系下的条件熵推广到邻域关系下,定义了邻域决策系统的条件熵。

定义 5 NDT = $\langle U,C\cup D,V,f\rangle$ 是一个邻域决策系统,对于 $\forall B\subseteq C$,对象 x 在属性集 B 下的邻域为 $\delta_B(x)$,D 在 U 上导出的划分为 $\{Y_1,Y_2,\cdots,Y_m,\cdots\}$,则决策属性集 D 相对于 B 的邻域条件熵定义为

$$NH(D|B)=-\int_{x\in U}\sum_{i=1}^{\infty}\frac{m(\delta_B(x)\cap Y_i)}{m(U)}\log_2\frac{m(\delta_B(x)\cap Y_i)}{m(\delta_B(x))}\mathrm{d}x$$

2 基于邻域近似条件熵的启发式属性约简

2.1 邻域近似条件熵及其性质

邻域近似精度可以有效地度量边界域引起的集合的不精确性,邻域条件熵则可以有效地度量信息粒度引起的知识不确定性[19],这两种单一的度量模型存在一定的局限性。基于此,将文献[16]的模型推广到无限集合,定义基于无限集的新模型。

定义 6 给定一个邻域决策系统 NDT = $\langle U,C\cup D,V,f\rangle$,对于 $\forall B\subseteq C$,对象 x 在属性 B 下的邻域为 $\delta_B(x)$,$U/\mathrm{IND}(D)=\{Y_1,Y_2,\cdots,Y_m,\cdots\}$,$\alpha_B(Y_i)$ 为 Y_i 在邻域关系 N_B 下的邻域近似精度,则决策属性集 D 相对于 B 的邻域近似条件熵定义为

$$\mathrm{NAH}(D|B)=-\sum_{i=1}^{\infty}\log_2(2-\alpha_B(Y_i))\times\int_{x\in U}\left(\frac{m(\delta_B(x)\cap Y_i)}{m(U)}\log_2\frac{m(\delta_B(x)\cap Y_i)}{m(\delta_B(x))}\right)\mathrm{d}x$$

定理 1 给定一个邻域决策系统 NDT = $\langle U,C\cup D,V,f\rangle$,对于 $\forall B\subseteq C$,对象 x_i 在属性集 B 下的邻域为 $\delta_B(x_i)$,$U/\mathrm{IND}(D)=\{Y_1,Y_2,\cdots,Y_m,\cdots\}$。邻域近似条件熵满足如下性质:a)当且仅当 $\forall x_i\in U,\delta_B(x_i)=U$,$\forall Y_j\in U/\mathrm{IND}(D),|Y_j|=1$ 时,$\mathrm{NAH}(D|B)$ 取得最大值 $m(U)\log_2 m(U)$;b)当且仅当所有对象都是正域中的元素时,$\mathrm{NAH}(D|B)$ 取得最小值 0。

证明 a)如果 $\forall Y_j\in U/\mathrm{IND}(D),|Y_j|=1$,并且 $\forall x_i\in U,\delta_B(x_i)=U$,则由定义 4 可知,对于任意的 $i\geqslant 1,\alpha_B(\{x_i\})=0$,进一步得 $\log_2(2-\alpha_B(\{x_i\}))=1$。另外,由于

$$\frac{m(\delta_B(x_i)\cap Y_j)}{m(U)}\log_2\frac{m(\delta_B(x_i)\cap Y_j)}{m(\delta_B(x_i))}=\frac{1}{m(U)}\log_2\frac{1}{m(U)}$$

再由定义 6 可得 $\mathrm{NAH}(D|B)=m(U)\log_2 m(U)$。

b)如果所有对象都是正域中的元素,即 $\forall x_i\in U,\delta_B(x_i)\subseteq\delta_D(x_i)=[x_i]_D$,则对于任意的 $j\geqslant 1$,$U/\mathrm{IND}(D)$ 中的等价类 Y_j 为若干邻域信息粒的并集。因此由定义 4 可知,对于任意的 $j\geqslant 1,\alpha_B(Y_j)=1$,进一步得 $\log_2(2-\alpha_B(Y_j))=0$。再由定义 6 可得 $\mathrm{NAH}(D|B)=0$。

定理 2 给定一个邻域决策系统 NDT = $\langle U,C\cup D,V,f\rangle$,$P,Q\subseteq C$,如果 $Q\subseteq P$,那么 $\mathrm{NAH}(D|P)\leqslant\mathrm{NAH}(D|Q)$。

证明 如果 $Q\subseteq P\subseteq C$,对于 $\forall X\subseteq U$,由定义 2 可知,$\underline{N_Q}X\subseteq\underline{N_P}X,\overline{N_Q}X\supseteq\overline{N_P}X$,由定义 4 可以进一步得出 $\alpha_P(X)\geqslant\alpha_Q(X)$。另外,假设 $U/\mathrm{IND}(D)=\{Y_1,Y_2,\cdots,Y_m,\cdots\}$,由性质 1 可得

$$0\leqslant-\frac{m(\delta_P(x)\cap Y_i)}{m(U)}\log_2\frac{m(\delta_P(x)\cap Y_i)}{m(\delta_P(x))}\leqslant-\frac{m(\delta_Q(x)\cap Y_i)}{m(U)}\log_2\frac{m(\delta_Q(x)\cap Y_i)}{m(\delta_Q(x))}$$

结合定义 6 可得 $\mathrm{NAH}(D|P)\leqslant\mathrm{NAH}(D|Q)$。

定理 2 表明 $\mathrm{NAH}(D|B)$ 随着属性集 B 中元素个数的增加而单调减少,这对于构建前向贪心属性约简算法非常重要。

定义 7[16] 给定一个邻域决策表 NDT = $\langle U,C\cup D,V,f\rangle$,$B\subseteq C$,当且仅当 a)$\mathrm{NAH}(D|B)=\mathrm{NAH}(D|C)$;b)$\forall b\in B$,有 $\mathrm{NAH}(D|B-\{b\})>\mathrm{NAH}(D|C)$,称 B 是 C 相对于 D 的一个约简。

第一个条件保证了选择的属性子集与整个属性集具有同样的信息度量;第二个条件保证了约简集中没有冗余的特征。

定义 8 内部属性重要度。给定一个邻域决策表 NDT = $\langle U,C\cup D,V,f\rangle$,则对于任意属性 $a\in C$ 在 C 中相对于 D 的内部属性重要度定义为

$$\mathrm{SIG}(a,C,D)=\mathrm{NAH}(D|C-\{a\})-\mathrm{NAH}(D|C)$$

定义 9[16] 给定一个邻域决策表 NDT = $\langle U,C\cup D,V,f\rangle$,对于任意 $a\in C$,如果 $\mathrm{NAH}(D|C-\{a\})>\mathrm{NAH}(D|C)$,即 $\mathrm{SIG}(a,C,D)>0$,则称属性 a 为 C 相对于 D 的一个核属性。

定义 10[16] 外部属性重要度。给定一个邻域决策表 NDT = $\langle U,C\cup D,V,f\rangle$,$B\subseteq C$,则对于任意属性 $a\in C-B$,a 相

对于 D 的外部属性重要度定义为

$$\mathrm{SIG}(a,B,D)=\mathrm{NAH}(D|B)-\mathrm{NAH}(D|B\cup\{a\})$$

$\mathrm{SIG}(a,B,D)$ 表示增加属性 a 后对于条件属性集 B 重要度的变化程度,$\mathrm{SIG}(a,B,D)$ 越大,a 在 B 中相对于 D 越重要。

2.2　算法描述

基于 2.1 节提出的熵模型,根据邻域近似条件熵的单调性原理,以定义 8、10 的属性重要度为度量指标,本文可以构造一种新的邻域决策表启发式属性约简算法。该算法从获取邻域决策表的核属性开始,然后循环选择使 $\mathrm{NAH}(D|B\cup\{a\})$ 最小(即外部属性重要度最大)的条件属性 a 添加到上一轮的约简属性集中,直到满足终止条件 $\mathrm{NAH}(D|B)=\mathrm{NAH}(D|C)$。

邻域粗糙集模型下基于邻域近似条件熵的属性约简改进算法如下。

算法 1　新的基于邻域近似条件熵的启发式属性约简

输入:邻域决策系统 $\mathrm{NDT}=\langle U,C\cup D,V,f\rangle$ 和 $\delta\geqslant 0$。

输出:邻域决策系统的一个约简 B。

a)初始化 $B=\varnothing$;

b)计算 NDT 中决策属性 D 相对于 C 的邻域近似条件熵 $\mathrm{NAH}(D|C)$;

c)对于任意的 $a_i\in C$,计算 $\mathrm{SIG}(a_i,C,D)$,如果 $\mathrm{SIG}(a_i,C,D)>0$,则 $B=B\cup\{a_i\}$;

d)此时,若 $B\neq\varnothing$,计算 $\mathrm{NAH}(D|B)$,如果 $\mathrm{NAH}(D|B)=\mathrm{NAH}(D|C)$,则转到 f);否则,执行 e);若 $B=\varnothing$,直接执行 e);

e)令 $R=C-B$,并执行如下步骤:

(a)对于每一个 $a\in R$,计算 $\mathrm{SIG}(a,B,D)$;

(b)选择 a_i,使其满足 $\mathrm{SIG}(a_i,B,D)=\max\{a|\mathrm{SIG}(a,B,D),a\in R\}$(若满足条件的属性有多个,则随机选择一个属性);

(c)令 $R=R-\{a_i\}$,$B=B\cup\{a_i\}$,计算新的 $\mathrm{NAH}(D|B)$,然后判断 $\mathrm{NAH}(D|B)=\mathrm{NAH}(D|C)$ 是否成立,若成立,执行 f)。若不成立,转到 a);

f)返回约简结果 B,算法结束。

2.3　算法时间复杂度分析

设 $|U|$ 和 $|C|$ 分别表示邻域决策系统中的样本数和条件属性个数。步骤 b)计算 $\mathrm{NAH}(D|C)$ 的时间复杂度为 $O(|C||U|^2)$;步骤 c)计算一次 $\mathrm{SIG}(a_i,C,D)$ 的时间复杂度为 $O(|C||U|^2)$,需要计算 $|C|$ 次,所以计算 core 的时间复杂度为 $O(|C|^2|U|^2)$;步骤 e)中,计算单个属性重要度的时间复杂度为 $O(|C||U|^2)$。在最坏的情况下,$|C|$ 可以取 1,2,3,…,$|C|-1$,$|C|$,所以步骤 e)的时间复杂度为 $O(|U|^2|C|^3)$。因此,该算法总的时间复杂度为 $O(|U|^2|C|^3)$。

3　实例分析

给定邻域决策表(表 1)[20] $\mathrm{NDT}=\langle U,C\cup D,V,f\rangle$,其中 $U=\{x_1,x_2,\cdots,x_{10}\}$,$C=\{a_1,a_2,a_3,a_4,a_5\}$;$D$ 为决策属性,取值为 $\{1,2,3\}$。通过邻域决策表 1 验证所提算法的可行性。

表 1　邻域决策表

U	a_1	a_2	a_3	a_4	a_5	D
x_1	0.21	0.63	0.31	0.50	0.44	1
x_2	0.65	0.49	0.79	0.15	0.64	1
x_3	0.48	0.21	0.36	0.22	0.63	1
x_4	0.51	0.21	0.31	0.48	0.52	2
x_5	0.46	0.76	0.61	0.45	0.38	2
x_6	0.51	0.42	0.28	0.16	0.56	1
x_7	0.37	0.22	0.42	0.52	0.69	3
x_8	0.35	0.82	0.06	0.19	0.41	1
x_9	0.31	0.61	0.74	0.16	0.40	1
x_{10}	0.46	0.59	0.19	0.26	0.83	3

对于表 1,使用算法 1 进行属性约简,给定 $\delta=0.2$,约简步骤如下:

a)根据定义 6 计算 $\mathrm{NAH}(D|C)=0$。

b)对于任意的 $a_i\in C$,计算 $\mathrm{SIG}(a_i,C,D)$,分别得出:

$$\mathrm{SIG}(a_1,C,D)=0,\mathrm{SIG}(a_2,C,D)=0$$

$$\mathrm{SIG}(a_3,C,D)=0,\mathrm{SIG}(a_4,C,D)=0.2199,\mathrm{SIG}(a_5,C,D)=0$$

因为 $\mathrm{SIG}(a_4,C,D)>0$,所以 $B=B\cup\{a_4\}=\{a_4\}$。

c)此时 $B\neq\varnothing$,计算 $\mathrm{NAH}(D|B)$。$\mathrm{NAH}(D|B)=5.3152\neq\mathrm{NAH}(D|C)$,继续执行。

d)$R=C-B=\{a_1,a_2,a_3,a_5\}$,对于每一个 $a\in R$,计算 $\mathrm{SIG}(a,B,D)$,分别得出:

$$\mathrm{SIG}(a_1,B,D)=2.2533,\mathrm{SIG}(a_2,B,D)=4.1669$$

$$\mathrm{SIG}(a_3,B,D)=3.4423,\mathrm{SIG}(a_5,B,D)=4.7458$$

选择最大的 a_i,使其满足 $\mathrm{SIG}(a_i,B,D)=\max\{a|\mathrm{SIG}(a,B,D),a\in R\}$,所以选择 a_5,此时,$B=B\cup\{a_i\}=\{a_4,a_5\}$。

计算新的 $\mathrm{NAH}(D|B)$,有 $\mathrm{NAH}(D|B)=0.5694\neq\mathrm{NAH}(D|C)$,继续执行。

e)此时 $R=R-\{a_5\}=\{a_1,a_2,a_3\}$,对于任意的 $a\in R$,计算 $\mathrm{SIG}(a,B,D)$,计算得出:

$$\mathrm{SIG}(a_1,B,D)=0.5694,\mathrm{SIG}(a_2,B,D)=0.5694$$

$$\mathrm{SIG}(a_3,B,D)=0.3495$$

选择最大的 a_i,使其满足 $\mathrm{SIG}(a_i,B,D)=\max\{a|\mathrm{SIG}(a,B,D),a\in R\}$,选择 a_1 或者 a_2,得到 $B=B\cup\{a_i\}=\{a_4,a_5,a_1\}$ 或 $B=B\cup\{a_i\}=\{a_4,a_5,a_2\}$,计算新的 $\mathrm{NAH}(D|B)$ 可得 $\mathrm{NAH}(D|B)=0=\mathrm{NAH}(D|C)$。因此最终约简结果为 $B=\{a_4,a_5,a_1\}$ 或者 $B=\{a_4,a_5,a_2\}$,结束。

4　实验验证

为了验证算法的有效性,下面将本文算法与代数观和信息观下的具有代表性的邻域粗糙集属性约简算法从属性约简个数和分类精度两方面进行比较。由于参考文献[9~11]都是代数观下的属性约简算法,实验结果其实是差不多的,所以实验只选文献[9]中的算法;同理信息观下的属性约简算法只选文献[18]中的算法;实验环境:ThinkPad E420 笔记本电脑,Intel Core i3-2330M CPU,4 GB 内存,Windows 7 的 32 位操作系统,采用 MATLAB R2010b 编程软件实现。本文选取 UCI 数据库的五个有限数据集作为实验数据集,数据集的描述如表 2 所示。

表 2　UCI 数据集描述

序号	数据集	样本数	条件属性数	决策类别数
1	wdbc	569	31	2
2	heart	270	13	7
3	sonar	208	60	2
4	ionosphere	351	34	2
5	vehicle	846	18	4

实验中,算法的距离函数采用 Manhattan 距离。为了减少不同属性量纲对约简结果的影响,计算样本邻域时,所有的数值型属性都被标准化到[0,1]内,本文采用最大最小值归一化方法,计算公式为

$$f(x_i)=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}}$$

同时,实验设置 $\delta=0.15$(邻域参数 $\delta=0.15$ 通过实验得到)。为了比较约简结果的分类能力,引入 SVM 分类器,以十

折交叉验证法进行实验,计算约简属性集的分类精度。其中分类精度的计算公式为

分类精度 = 被正确分类的样本数/样本总数 ×100%[11]

具体实验结果如表 3 ~5 所示。

表3 不同算法约简个数比较

数据集	约简后条件属性个数		
	文献[9]的算法	文献[18]的算法	本文算法
wdbc	6	6	6
heart	8	8	7
sonar	5	5	5
ionosphere	6	6	6
vehicle	9	8	8

表4 不同算法约简得到的属性

数据集	约简后得到的条件属性		
	文献[9]的算法	文献[18]的算法	本文算法
wdbc	24,29,3,26,13,23	29,23,2,30,26,20	29,23,2,30,26,20
heart	7,10,12,3,4,1,8,5	3,12,11,1,4,8,7,5	3,12,1,8,4,7,11
sonar	1,45,37,21,30	21,36,30,15,1	21,36,30,15,11
ionosphere	1,5,17,24,9,4	21,24,20,3,4,5	21,22,25,14,3,6
vehicle	9,18,15,16,10,13,4,2,17	2,18,15,16,1,7,10,5	2,18,15,16,1,7,17,5

表5 SVM 分类器下的分类精度 /%

数据集	约简子集的分类精度		
	文献[9]的算法	文献[18]的算法	本文算法
wdbc	96.66	96.66	96.66
heart	77.78	78.89	82.59
sonar	71.63	72.60	75.48
ionosphere	91.74	90.60	93.16
vehicle	77.31	74.11	78.25

表3 和4 分别展示了三种算法约简后的条件属性个数和得到的具体条件属性。从两个表可以看出:a)本文提出的约简算法同现有的代数观和信息观下的属性约简算法一样,都能大量地减少条件属性数量,有效地实现了对数据集的约简,这进一步说明了本文算法是有效的,为属性约简提供了一种新的思路;b)本文算法在各个数据集上约简后的条件属性个数与其他两种类型的约简算法基本相同,在一些数据集能够获得更少的属性,结果较紧凑。如 Heart 数据集在基于正域的约简算法和基于条件熵的约简算法下得到的约简数量为八个,而经过本文所提属性约简算法进行约简后数量为七个。

表5 展示了三种算法约简后的分类精度。从表 5 的实验对比结果可以看出:a)在属性约简个数相同的数据集上,由于约简所得到的条件属性存在差异,分类精度也是不同的;b)本文算法在获得较紧凑约简的同时,约简子集的分类精度优于或等于其他两种类型的算法。除了在 wdbc 上等于其他算法外,在其他数据集上,本文所提出的算法具有较高的精度。这是由于邻域近似条件熵结合了邻域近似精度和邻域条件熵,是一种更加完备的属性重要性度量模型。

5 结束语

邻域粗糙集模型能够直接处理数值型信息系统,其扩展模型和属性约简是邻域粗糙集模型的主要研究内容。本文给出无限集下邻域近似条件熵的定义,探讨了其基本性质,邻域近似条件熵结合了邻域粗糙集的代数观点和信息观点。根据邻域近似条件熵的单调性原理,设计了一种新的基于邻域近似条件熵的属性约简算法,并与代数观和信息观下的属性约简算法进行了比较实验。在 UCI 数据集上的实验验证了本文算法的有效性,实验结果表明,与现有的邻域粗糙集属性约简算法相比,本文所提出的约简算法不仅能够获得紧凑的约简,而且具有较好的分类性能。由于在实际应用中,信息系统常常是动态变化的,为此下一步工作将研究邻域粗糙集的动态属性约简算法。此外,降低算法的时间复杂度,寻求更加高效的约简算法也是本文下一步研究的重点。

参考文献:

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

[2] 周建华,徐章艳,章晨光. 改进的差别矩阵的快速属性约简算法[J]. 小型微型计算机系统,2014,35(4):831-834.

[3] 杨传健,葛浩,汪志圣. 基于粗糙集的属性约简方法研究综述[J]. 计算机应用研究,2012,29(1):16-20.

[4] Qian Yuhua, Liang Jiye, Pedrycz W, *et al*. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence,2010,174(9):597-618.

[5] Liang Jiye, Mi Junrong, Wei Wei, *et al*. An accelerator for attribute reduction based on perspective of objects and attributes[J]. Knowledge-Based Systems,2013,44(1):90-100.

[6] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001:99-116.

[7] Lin T Y. Rough sets, neighborhood systems and approximation[J]. World Journal of Surgery,1986,10(2):189-194.

[8] Wu Weizhi, Zhang Wenxiu. Neighborhood operator systems and approximations[J]. Information Sciences,2002,144(1-4):201-217.

[9] 胡清华,赵辉,于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能,2008,21(6):730-738.

[10] Hu Qinghua, Yu Daren, Liu Jinfu, *et al*. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences,2008,178(18):3577-3594.

[11] 梁海龙,谢珺,续欣莹,等. 新的基于区分对象集的邻域粗糙集属性约简算法[J]. 计算机应用,2015,35(8):2366-2370.

[12] Hu Qinghua, Zhanglei, Zhang D, *et al*. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert Systems with Applications, 2011, 38(9): 10737-10750.

[13] 续欣莹,刘海涛,谢珺,等. 信息观下基于不一致邻域矩阵的属性约简[J]. 控制与决策,2016,31(1):130-136.

[14] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766.

[15] 江峰,王莎莎,杜军威,等. 基于近似决策熵的属性约简[J]. 控制与决策,2015,30(1):65-70.

[16] 谢玲玲,雷景生,徐菲菲. 基于改进的邻域粗糙集与概率神经网络的水电机组振动故障诊断[J]. 上海电力学院学报,2016,32(2):181-187.

[17] Halmos P R. Measure theory[M]. [S.l.]:World Publishing Corporation, 2007: 100-152.

[18] Dai Jianhua, Wang Wentao, Xu Qing. An uncertainty measure for incomplete decision tables and its applications[J]. IEEE Trans on Cybernetics,2013,43(4):1277-1289.

[19] 滕书华,鲁敏,杨阿锋,等. 基于一般二元关系的粗糙集加权不确定性度量[J]. 计算机学报,2014,37(3):649-665.

[20] 唐朝辉,陈玉明. 邻域系统的不确定性度量方法[J]. 控制与决策,2014,29(4):691-695.