西北大学学报(自然科学版)
2022 年 10 月,第 52 卷第 5 期,Oct. ,2022,Vol. 52,No. 5
Journal of Northwest University (Natural Science Edition)

·不确定性知识发现·

可变精度邻域区间值决策表的属性约简

徐伟华,李思琪

(西南大学 人工智能学院,重庆 400715)

摘要:区间值决策表可以通过区间刻画对象相对于条件属性的取值,其在现实生活中运用非常广泛,该文在此基础上提出一种启发式约简算法。首先,为了提高信息系统的可辨识性,在区间值决策系统上引入可变精度阈值 α 与邻域阈值 δ;随后,重新定义了决策表的上近似、下近似与正域;最后,定义属性质量度,以判定是否将条件纳入约简集合。为了更好地理解该算法的执行过程,该文进行了案例分析,并在 4 个数据集上完成了实验。实验结果表明,该算法具有良好的可行性,且其在准确率上优于另 2 种算法。

关键词:区间值决策表;可变精度邻域决策表;属性约简;正域;属性质量度

中图分类号:TP189 **DOI**:10.16152/j.cnki.xdxbzr.2022-05-001

Attribute reduction of interval-valued neighborhood decision table with variable precision

XU Weihua, LI Siqi

(College of Artificial Intelligence, Southwest University, Chongqing 400715, China)

Abstract: Interval valued decision tables are widely used for various fields in reality. This paper proposes a heuristic reduction algorithm for this kind of decision tables. Firstly, the variable accuracy threshold α and the neighborhood threshold δ are defined on the interval-valued decision table. Then, this paper redefines the upper approximation, lower approximation and positive domain of decision table. Finally, attribute quality is defined to determine whether conditions should be included in the reduction set. In order to better understand the execution process of the algorithm, this paper conducts a case analysis and completes the experiment on four datasets. Experimental results show that the proposed algorithm has good feasibility, and its accuracy is better than the other two algorithms.

Key words: interval-valued decision table; variable precision neighborhood decision table; attribute reduction; positive domain; attribute quality

1965 年,Zadeh 首次提出了模糊集的概念[1],标志着模糊数学的诞生。该学科在 1995 年被 ACM 列为新兴的计算机科学研究领域,如今正在继续发展。因其建立在分类基础上,可以有效处理不完整不确定问题,所以在实践中广泛应用[2]。同时,区间值决策表[3-4]作为一个分支,能很好描绘不精确对象的特征,在医学、金融、机械制造等领域意义重大。Lin 和 Hu 在 Zadeh 的知

收稿日期:2022-03-29

基金项目:国家自然科学基金项目(61976245)

第一作者:徐伟华,男,山西浑源人,博士,教授,博士生导师,从事粒计算与知识发现、机器学习、不确定性人工智能、概念认知学习等研究,E-mail:chxuwh@swu.edu.cn。

识粒化的基础上将邻域引入粗糙集,以粗糙集理论为基础,衍生出了邻域粗糙集理论。该理论重新定义上下近似,实现了一种全新的近似逼近。邻域粗糙集理论已经广泛应用在决策分析、过程控制以及模式识别等[5-9]领域。

在使用过程中需要对属性值进行属性约简。属性约简是粗糙集理论研究的核心问题之一,决策表中有一些条件属性,由于其属性值难以测量或测量这些属性值花费极高,需要将之删去。在保持分类水平不变的情况下,尽力删除这些冗余属性,使剩余属性达到最简,以降低统计难度。这就是属性约简。事实上,寻找约简集合[10-13]是 NP-hard 问题,解决这类问题一般是采用启发式搜索以获得近似解。

然而,属性约简后的属性值不可避免会丢失部分原始数据,导致一定程度的信息缺失,限制了粗糙集的应用范围,为了解决这个问题,前人用邻域关系代替等价关系,重新定义上下近似与正域,建立了可变精度邻域决策表[14-15]及相应的属性约简算法。但是现实生活中,区间值决策表应用范围更加广泛。如果能将这种方法推广到区间值决策表,会得到更广泛的应用。

为此,本文将经典可变精度邻域信息决策表推广到区间值信息决策表上,定义区间距离以用于计算邻域,用可变精度阈值计算出条件的正域,对信息表进行属性约简。并以属性质量度为判断依据,设计相应启发式属性约简算法。最后,通过实验验证了算法的正确性。

1 基本概念

本节将介绍可变精度邻域决策表以及区间值信息决策表的相关概念。

1.1 可变精度邻域决策表

一个决策表[2]可表示为二元组 DT = $\langle U, \mathrm{AT} \cup d\rangle$,其中非空有限集合 $U = \{x_1, x_2, \cdots, x_n\}$ 为对象集,称为全域或样本空间,AT 表示一个非空的有限条件属性集合,用于描述 U 的实数型特征,d 是决策属性。

$\forall x \in U, \forall a \in \mathrm{AT}, a(x)$ 表示样本 x 在属性 a 上的取值,而 $d(x)$ 为样本 x 在决策属性 d 上的值,$U/d = \{X_1, X_2, \cdots, X_m\}$ 代表 U 被决策属性 d 划分出的决策类。

给定一个决策表 DT = $\langle U, \mathrm{AT} \cup d\rangle$,且邻域半径 $\delta \in (0,1)$,则对于 $\forall x \in U$,邻域 $\delta(x)$ 定义为

$$\delta(x) = \{x_j \mid x_j \in U, \Delta(x, x_j) \leqslant \delta, \delta > 0\},$$

其中,$\Delta(x, x_j)$ 代表对象 x 和 x_j 之间的距离。

给定一个决策表 DT = $\langle U, \mathrm{AT} \cup d\rangle$,设集合 $X \subseteq U$,集合 $Y \subseteq U$,则 X 关于 Y 的错误分类率可表示为

$$e_X^Y = \min\left(1, \max\left\{1 - \frac{|X \cap Y|}{|X|}, 0\right\}\right)。$$

定义 1 给定一个邻域信息决策表 NDT = $\langle U, \mathrm{AT} \cup d\rangle$,该决策表中有 m 个等价类 $U/d = \{X_1, X_2, \cdots, X_m\}$。对于 $\forall B \subseteq \mathrm{AT}$,引入可变精度的正确率阈值 $\alpha(0.5 \leqslant \alpha \leqslant 1)$。则该精度下邻域信息决策表相对于决策属性 d 的上近似为

$$\overline{P}_B^{\alpha}(d) = \bigcup_{i=1}^{m} \overline{P}_B^{\alpha}(X_i);$$

下近似为

$$\underline{P}_B^{\alpha}(d) = \bigcup_{i=1}^{m} \underline{P}_B^{\alpha}(X_i);$$

正域为

$$\underline{P}_B^{\alpha}(d) = \mathrm{Pos}_B(d),$$

其中:

$$\overline{P}_B^{\alpha}(X_i) = \{x \mid e_{\delta(x)}^{X_i} < 1 - \alpha, x \in U\};$$

$$\underline{P}_B^{\alpha}(X_i) = \{x \mid e_{\delta(x)}^{X_i} \leqslant \alpha, x \in U\}。$$

传统意义的邻域决策表在定义上下近似时并未考虑容错率,因此对错误的分类非常敏感。为了更好地处理不确定关系以及减少噪声干扰,更常使用具有一定容错性的可变精度邻域决策表。

如上文所示,可变精度邻域粗糙集的上下近似是基于 α 的容错划分,通过增大 α 的值,使之具有更好的覆盖率。α 越小,正域将扩大,容错率也变大;相反的,α 越大,正域越小,容错率越小,上下近似越精确。我们需要选择一个合适的 α,使决策表具有良好辨识性的同时,保证一定容错率。

1.2 区间值信息决策表

一个区间值决策表可以表示为 IVDT = $\langle U, \mathrm{AT} \cup d, V_{\mathrm{AT}}, f\rangle$,其中,决策属性 d 的取值同经典情况(即单值,而非区间值)。V_a 为任意条件属性 $a \in v\mathrm{AT}$ 的值域,那么其条件属性值域 $V_{\mathrm{AT}} = \cup_{a \in \mathrm{AT}} V_a$。信息函数 $f: U \times \mathrm{AT} \to V_{\mathrm{AT}}$ 满足 $\forall x_i \in U, \forall a \in \mathrm{AT}$,且 $f(x_i, a)$ 为一个区间值。

定义 2 设有两个不同的区间 A 与 B,区间 $A = [a^-, a^+], B = [b^-, b^+]$。则 A 区间相对于 B 区间的优势度定义为

$P_{A\geqslant B}=\min(1,\max\{\frac{a^+-b^-}{(a^+-a^-)+(b^+-b^-)},0\})$。由该定义易知,

1) $P_{A\geqslant B}\neq P_{B\geqslant A}$;

2) $0\leqslant P_{A\geqslant B}\leqslant 1$;

3) $P_{A\geqslant B}+P_{B\geqslant A}=1$;

4) $P_{A\geqslant A}=0.5$。

进一步给出区间值决策表下 2 个对象的距离。设有区间值决策表 IVDT = $\langle U,\mathrm{AT}\cup d\rangle$,其中,$x$ 与 y 是全域 U 中的2个不等的对象,属性子集 B 中一共有 s 个条件。在条件 k 下,x 与 y 对应的取值区间分别是 A_i^k 和 A_j^k。则 x 与 y 对应的取值区间分别是 A_i^k 和 A_j^k。则 x 与 y 的距离可以表示为

$$\Delta(x,y)=\sqrt{\sum_{k=1}^{s}(P_{A_i^k\geqslant A_j^k}-P_{A_j^k\geqslant A_i^k})^2}。$$

即2个对象在条件集 AT 下的欧氏距离,通过这个关系,可以将区间值决策表与邻域可变精度决策表结合到一起。显然,它满足如下关系,

1) $\Delta(x,x)=0$;

2) $\Delta(x,y)=\Delta(y,x)$;

3) $\Delta(x,z)\leqslant\Delta(x,y)+\Delta(y,z)$。

2　变精度邻域区间决策表属性约简

本节将可变精度邻域粗糙集引入区间值信息决策表,并提出该决策表的约简方式。

2.1　属性质量度

给定一个区间值邻域决策表 INDT = $\langle U,\mathrm{AT}\cup d\rangle$,其中,$\forall B\subseteq\mathrm{AT},\forall a\in\mathrm{AT}-B$,且 X 是由决策属性 d 划分而出的等价类。现有 $x_i\in\mathrm{Pos}_{B\cup\{a\}}(d)$,$x_j\in\mathrm{Pos}_B(d)$,则属性 a 相对于属性子集 B 的平均正确分类率的增量函数定义为

$$\mathrm{Inc}_a^B=\frac{\sum(1-e^X_{\delta(x_i)})}{|\mathrm{Pos}_{B\cup d\{a\}}(d)|}-\frac{\sum(1-e^X_{\delta(x_j)})}{|\mathrm{Pos}_b(d)|}。$$

即正域改变前后其内所有样本正确分类率求和取均值后的增量,显然有

1) 当 $\mathrm{Inc}_a^B<0$,所添加属性 a 使得正确分类率降低;

2) 当 $\mathrm{Inc}_a^B=0$,所添加属性 a 没有使正确分类率发生变化;

3) 当 $\mathrm{Inc}_a^B>0$,所添加属性 a 使得正确分类率升高。

区间值邻域决策表 INDT = $\langle U,\mathrm{AT}\cup d\rangle$,$B\subseteq\mathrm{AT}$,若 $a\in\mathrm{AT}-B$,则 a 相对于属性子集 B 关于决策属性 d 的正域增量函数,可用正域与全域基数之比的增量表示,即文献[13]中定义的属性重要性,其定义为

$$\mathrm{Sig}_a^B=\frac{|\mathrm{Pos}_{B\cup\{a\}}(d)|-|\mathrm{Pos}_B(d)|}{|U|}。$$

即增加属性 a 前后正域的相对改变量,显然有

1) 当 $\mathrm{Sig}_a^B<0$,所添加属性 a 使得正域变小;

2) 当 $\mathrm{Sig}_a^B=0$,所添加属性 a 没有使正域发生变化;

3) 当 $\mathrm{Sig}_a^B>0$,所添加属性 a 使得正域变大。

定义 4　区间值邻域决策表 INDT = $\langle U,\mathrm{AT}\cup d\rangle$,若 $\forall B\subseteq\mathrm{AT},\forall a\in\mathrm{AT}-B$,则 a 相对于属性子集 B 关于决策属性 d 的属性质量度可以定义如下,

$Q_a^B=\mathrm{Inc}_a^B\times\mathrm{Sig}_a^B$。

属性质量度 Q_a^B 是增量值的积,反映了属性 a 是否对系统正确分类率与正域有所贡献,可以最大程度地反应一个区间值属性对全局约简的重要性。显然,它也具有以下性质,

1) $Q_a^B\geqslant 0$,即任意属性的质量度都不为负;

2) Q_a^B 越大,说明属性 a 越重要;

3) $Q_a^B=0$,说明增加属性 a 后正确分类率没变或者正域没变,也就说明 a 是冗余的,可以被约简。

2.2　属性约简

给定区间值邻域决策表 INDT = $\langle U,\mathrm{AT}\cup d\rangle$,若 red 是 AT 的一个约简集合,对于 $\forall a\in\mathrm{red}$,red 需满足

1) $\mathrm{Pos}_{\mathrm{red}-\{a\}}(d)<\mathrm{Pos}_{\mathrm{red}}(d)$;

2) $\mathrm{Pos}_{\mathrm{red}}(d)=\mathrm{Pos}_{\mathrm{AT}}(d)$。

属性质量度函数是正域增量与正确分类率增量的乘积,因此可以用属性质量度表示这种正域的变化。即,若 $\forall b\in\mathrm{AT}-\mathrm{red}$,以上关系也可表示为

1) $Q_a^{\mathrm{red}-\{a\}}>0$;

2) $Q_b^{\mathrm{red}}=0$。

即 red 中任意一个属性都是必不可少的,而 red 以外任意一个属性对 red 都是冗余的。这与经典集的定义是几乎一致的,只是增加了数值粒化而已。这样可以保证约简 red 与全部条件属性具有相同的分辨能力的同时达到最精简。

给定区间值邻域决表 INDT = $\langle U,\mathrm{AT}\cup d\rangle$,若 $B_1,B_2,\cdots,B_n$ 是该表的全部约简集合,则称 $\cap_{i\leqslant n}B_i$ 为此信息决策表的核。

3 案例分析

为了说明上一节属性约简的具体机理,本节给出一个具体案例以进行详细分析。

现从一个信息表中中抽取 8 个数据组成一个小型区间值信息决策表 INDT = $\langle U, \mathrm{AT} \cup d \rangle$, $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, 决策属性 $d = \{\mathrm{Rainfall}\}$ (Y 代表降雨,N 代表未降雨)。条件集 AT = {Vegetation, Humidity, Airflow Rainfall}。为方便后续计算,以首字母简写代替。并对其进行归一化,将区间值映射到[0,1],处理后的信息决策表见表 1。

若选取邻域为 $\delta = 0.3$,正确率阈值 $\alpha = 0.8$,以此表为实例进行计算。依次计算所有属性子集的邻域,如表 2 所示。

根据表 1,决策属性 Rainfall 将论域划分为 2 个等价类:$X_1 = \{x_1, x_4, x_5, x_6, x_8\}$, $X_2 = \{x_2, x_3, x_7\}$,初始化约简集合 red = ∅。

根据前文的定义,首先分别求得 3 个条件及条件全集的正域,

$\mathrm{Pos}_{\{V\}}(R) = \{x_1, x_3, x_4, x_5, x_6, x_7, x_8\}$,

$\mathrm{Pos}_{\{H\}}(R) = \{x_1, x_3, x_4, x_5, x_7, x_8\}$,

$\mathrm{Pos}_{\{A\}}(R) = \{x_1, x_3, x_4, x_5, x_6, x_7, x_8\}$,

$\mathrm{Pos}_{\{\mathrm{AT}\}}(R) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$。

表 1 关于降雨的影响因素的信息决策表

Tab. 1 An interval-valued information tableon analysis of factors affecting rainfall

U	Vegetation(V)	Humidity(H)	Airflow(A)	Rainfall(R)
x_1	[0.5, 0.8]	[0.2, 0.4]	[0.6, 0.9]	Y
x_2	[0.4, 0.7]	[0.2, 0.3]	[0.5, 0.8]	N
x_3	[0.2, 0.4]	[0.4, 0.5]	[0.3, 0.5]	N
x_4	[0.5, 0.7]	[0.3, 0.4]	[0.6, 0.8]	Y
x_5	[0.5, 0.7]	[0.2, 0.4]	[0.6, 0.8]	Y
x_6	[0.5, 0.8]	[0.2, 0.3]	[0.6, 0.9]	Y
x_7	[0.2, 0.5]	[0.3, 0.5]	[0.3, 0.6]	N
x_8	[0.5, 0.7]	[0.3, 0.4]	[0.6, 0.8]	Y

表 2 所有属性子集的邻域

Tab. 2 Neighborhood of all subsets

	$\{V\}$	$\{H\}$	$\{A\}$	$\{V,H\}$	$\{V,A\}$	$\{H,A\}$	$\{V,H,A\}$
$\delta(x_1)$	$x_{1,4,5,6,8}$	$x_{1,5}$	$x_{1,4,5,6,8}$	$x_{1,5}$	$x_{1,4,5,6,8}$	$x_{1,5}$	$x_{1,5}$
$\delta(x_2)$	$x_{2,4,5,8}$	$x_{2,6}$	$x_{2,4,5,8}$	x_2	$x_{2,4,5,8}$	x_2	x_2
$\delta(x_3)$	$x_{3,7}$	x_3	$x_{3,7}$	x_3	$x_{3,7}$	x_3	x_3
$\delta(x_4)$	$x_{1,2,4,5,6,8}$	$x_{4,8}$	$x_{1,2,4,5,6,8}$	$x_{4,8}$	$x_{1,2,4,5,6,8}$	$x_{4,8}$	$x_{4,8}$
$\delta(x_5)$	$x_{1,2,4,5,6,8}$	$x_{1,5}$	$x_{1,2,4,5,6,8}$	$x_{1,5}$	$x_{1,2,4,5,6,8}$	$x_{1,5}$	$x_{1,5}$
$\delta(x_6)$	$x_{1,4,5,6,8}$	$x_{2,6}$	$x_{1,4,5,6,8}$	x_6	$x_{1,4,5,6,8}$	x_6	x_6
$\delta(x_7)$	$x_{3,7}$	x_7	$x_{3,7}$	x_7	$x_{3,7}$	x_7	x_7
$\delta(x_8)$	$x_{1,2,4,5,6,8}$	$x_{4,8}$	$x_{1,2,4,5,6,8}$	$x_{4,8}$	$x_{1,2,4,5,6,8}$	$x_{4,8}$	$x_{4,8}$

进一步可求出 3 个条件的属性质量度,

$$Q_V^{\mathrm{red}} = \frac{7}{8} \times \frac{1 \times 4 + \frac{5}{6} \times 3}{7} = 0.812\,5,$$

$$Q_H^{\mathrm{red}} = \frac{6}{8} \times \frac{1 \times 6}{6} = 0.75,$$

$$Q_A^{\mathrm{red}} = \frac{7}{8} \times \frac{1 \times 4 + \frac{5}{6} \times 3}{7} = 0.812\,5,$$

根据计算结果,选取属性质量度最高的条件 V 或 A,即 $\mathrm{red}_1 = \{V\}$, $\mathrm{red}_2 = \{A\}$。

如果选取 red_1 为约简集合,再分别计算{V,

$H\}$、$\{V,A\}$ 的正域,

$\mathrm{Pos}_{\{V,H\}}(R) = \{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$;

$\mathrm{Pos}_{\{V,A\}}(R) = \{x_1,x_3,x_4,x_5,x_6,x_7,x_8\}$。

再分别计算条件 H 与条件 A 相对与 red_1 的属性质量度,

$$Q_H^{\mathrm{red}_1} = \frac{8-7}{8}\times\left(\frac{1\times 8}{8}-\frac{1\times 4+\frac{5}{6}\times 3}{7}\right)=\frac{1}{112};$$

$$Q_A^{\mathrm{red}_1} = \frac{7-7}{8}\times\left(\frac{1\times 4+\frac{5}{6}\times 3}{7}-\frac{1\times 4+\frac{5}{6}\times 3}{7}\right)=0。$$

说明条件 A 对于 red_1 是冗余的,选取属性质量度最大的条件 H 将其加入到约简集合 red_1。

又因 $\mathrm{Pos}_{\{V,H\}}(R) = \mathrm{Pos}_{\mathrm{AT}}(R)$,即正域不再发生变化,所以 red_1 = {Vegetation,Humidity} 即为约简集合。

同理,可求出 red_2 = {Humidity,Airflow} 也是一个约简集合,2 个约简集合的交集 {Humidity} 为该信息表的核。结果如表 3 所示。

表 3　约简集合及核

Tab. 3　The reduction of collection and core

red_1	red_2	核
{Vegetation,Humidity}	{Humidity,Airflow}	{Humidity}

4　算法设计与数值实验

4.1　算法设计及时间复杂度分析

具体算法如算法 1 所示。

算法 1　关于可变精度邻域区间值决策表属性约简的启发式算法

输入　区间邻域决策表 IVDT = $\langle U,\mathrm{AT}\cup d\rangle$,可变精度阈值 α,邻域取值 δ。

输出　属性约简集合 red。

```
1) begin
2)   compute U/d = {X_1, X_2, ..., X_m};
3)   red←∅; Q_max←0; /* 初始化约简集合和属性质量度 */
4)   for a ∈ AT − red do
5)     for x ∈ U do
6)       compute δ(x); /* 计算全体对象在{a}∪red 下的邻域 */
7)     end
8)     compute Q_a^red; /* 计算属性质量度 */
9)     if Q_a^red > Q_max then
10)      Q_max←Q_a^red; a_max←a; /* 更新属性质量度最大的属性 */
12)    end
13)  end
14)  if Q_max > 0 then
15)    red←red ∪ {a_max}; /* 属性质量度最大的属性被加入约简集合 */
16)    goto 4;
17)  else
18)    return red;
19)  end
20) end
```

接下来分析该算法的时间复杂度。在该算法中,循环体主要应用于求解邻域与计算条件的属性质量度中。假设一共有 n 个条件,最后得到的约简集合中条件 m 个,在此时刻约简了 k 个条件。

计算邻域时,在每个条件下计算各对象的邻域需要循环 $(n-k)c\frac{|U|\times(|U|-1)}{2}$ 次,所以时间复杂度为 $O(n\times|U|^2)$。

计算属性质量度时,求出每个条件的正域需要循环 $(n-k)\times|U|$ 次,求出各条件的属性质量度需要循环 $(n-k)$ 次,这两个循环是线性关系,所以时间复杂度为 $O(n\times|U|)$。

将新属性添加至约简集合的循环需要经历 $(m+1)$ 次,时间复杂度为 $O(n)$。

综上,时间复杂度为 $O(n^2\times|U|^2)$。

4.2　实验数据与实验环境

为了验证算法的正确性,本次实验选用 UCI 库上的 4 个分类数据集。

首先将非数值型的特征值替换为数值型,对数据使用 Min-Max 归一化将值映射到[0,1]区间,以消除量纲影响,随后将其按照下列方法转换为区间值信息决策表,使用算法对其进行属性约简。

要将传统数值的数据集转换为区间值数据集,我们将单值 $a_t(x_i)$ 转换为区间值 $[u_i^t,v_i^t]$。采用文献[5]中的方式进行转换,

$u_i^t = a_t(x_i) - 2\delta_t^k$;

$v_i^t = a_t(x_i) + 2\delta_t^k$。

其中:$a_t(x_i)$ 是 $x_i \in U$ 在条件属性 a_t 下的取值;δ_t^k 是关于所有 x_j 的 $a_t(x_j)$ 在决策类 D_k 下的标准差,可用下式得到,

$$\delta_t^k = \sqrt{\frac{1}{|D_k|-1}\sum_{x_j \in D_k}(a_t(x_j) - \bar{a}_t^k)^2};$$

$$\bar{a}_t^k = \frac{\sum_{x_j \in D_k} a_t(x_j)}{|D_k|}。$$

此外,为验证算法有效性,我们对其约简前后的分类能力做了对比,按照 8∶2 的比例划分训练集和测试集,并且选用支持向量机(SVM)与梯度提升模型(GBDT)对其进行验证。选用的数据集信息如表 4 所示。

表 4 数据集描述

Tab. 4 Dataset description

序号	数据集	$\|U\|$	$\|AT\|$	$\|d\|$
1	Ionosphere	351	34	2
2	Car Evaluation	1 728	6	4
3	Solar Flare	1 066	12	6
4	Abalone	4 177	8	28

4.3 实验结果分析

实验研究了算法在不同邻域阈值 $\delta(0.4 \sim 0.6)$ 和不同可变精度阈值 $\alpha(0.6 \sim 0.9)$ 下得出的约简集合以及约简前后分类预测准确率的变化。比较分类精度变化需要对样本进行机器学习,此时选取的邻域和变精度为

$$\sigma = 0.5,\ \alpha = 0.7$$

以此判断约简集合是否可以近似代表整个系统的信息。

图 1 反应了约简前后的数据集在支持向量机(SVM)和梯度提升决策树(GBDT)下的分类准确率的变化。表 5 为在上述参数下约简前后的准确率。实验结果表明,约简后的分类准确率均不小于约简前的准确率。说明算法选择的属性可以有效地近似数据集的分类能力。

表 6 为 4 个数据集使用本文的方法得出的约简集,其中的元素是决策属性的序号,可见在某些条件下约简集合不止一个。

同时,本文选 2 两种约简算法作为对比算法,分别是来自参考文献[3]中的 RDAR 算法和误分代价算法,比较了 3 种约简算法的准确率,结果如图 2 所示。可见本文算法在 4 个数据集上的准确率基本大于另外 2 种算法。

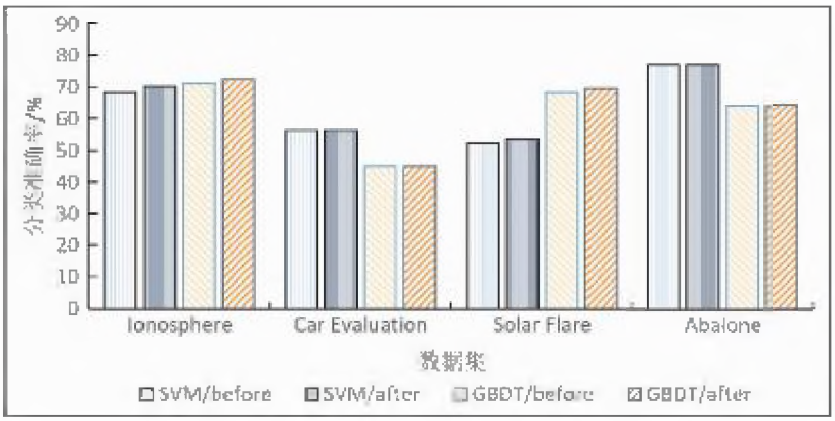

图 1 4 种数据集约简前后的准确率

Fig. 1 Classification accuracy before and after reduction of four data sets

表 5 一定条件下约简前后的分类精度

Tab. 5 Classification accuracy before/after reduction under certain conditions

模型	约简前/后	Ionosphere	Car Evaluation	Solar Flare	Abalone
SVM	约简前	68.23	56.44	52.32	77.25
	约简后	70.24	56.44	53.48	77.25
GBDT	约简前	71.30	44.98	68.21	64.35
	约简后	72.44	44.98	69.45	64.35

表 6　数据集的约简集结果

Tab. 6　Reduced set results for all datasets

δ	α	Ionosphere	Car Evaluation	Solar Flare	Abalone
0.4	0.6	{{5,6,7,9,24},{2,5,6}}	{1,2,3,4,5,6}	{1,5}	{{1,2,3,6,8},{1,2,3,5,8}}
	0.7	{{5,6,7,9,24},{2,5,6}}	{1,2,3,4,5,6}	{1,5}	{{1,2,3,6,8},{1,2,3,5,8}}
	0.8	{{5,6,7,9,24},{2,5,6}}	{1,2,3,4,5,6}	{1,5}	{{1,2,3,6,8},{1,2,3,5,8}}
	0.9	{{5,6,7,9,24},{2,5,6}}	{1,2,3,4,5,6}	{1,5}	{{1,2,3,6,8},{1,2,3,5,8}}
0.5	0.6	{{5,6,7,9,24},{2,5,6}}	{1,2,3,4,5,6}	{1,5}	{1,2,3,5,6,7,8}
	0.7	{{5,6,7,9,24},{2,5,6}}	{1,2,3,4,5,6}	{{1},{5}}	{1,2,3,5,6,7,8}
	0.8	{5,6,8,9,24,34}	{1,2,3,4,5,6}	{{1},{5}}	{1,2,3,5,6,7,8}
0.6	0.7	{5,6,8,9,24,34}	{{1,4,6},{2,4,6}}	{{1},{5}}	{1,2,3,5,6,7,8}
	0.8	{5,6,8,9,24,34}	{{1,4,6},{2,4,6}}	{{1},{5}}	{1,2,3,5,6,7,8}

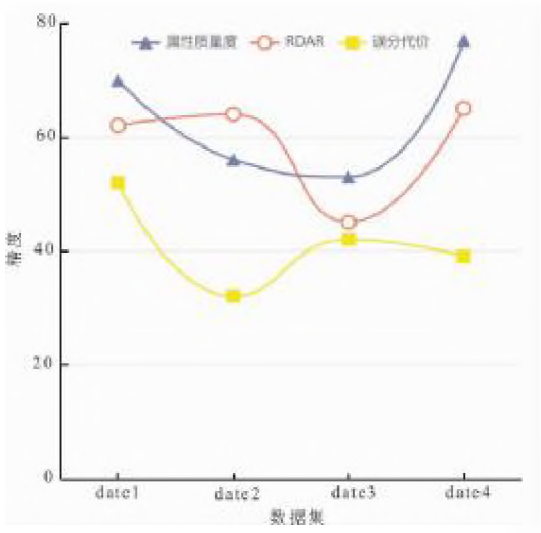

图 2　3 种约简算法准确率

Fig. 2　The accuracy of three reduction algorithms

5　结语

本文在基于可变精度邻域关系的区间值决策信息表的模型下,提出区间距离计算公式,并基于此提出该信息表中上下近似、核和正域的概念。同时,为了删除在数据采集过程中存在的一些不必要的条件属性,本文使用正域以及分类正确率的变化定义了属性质量度,设计了一种启发式属性约简算法,并通过实验验证了该算法的有效性。实验结果表明,该算法选择的属性可以近似原数据集的分类能力。

参考文献:

[1]　ZADEH L A. Fuzzy sets[J]. Information & Control, 1965, 8(3):338-353.

[2]　徐伟华. 序信息系统与粗糙集[M]. 北京:科学出版社, 2013.

[3]　DAI J H, WANG W T, MI J S. Uncertainty measurement for interval-valuedinformation systems[J]. Information Sciences, 2013, 251(4):63-78.

[4]　汪凌. 区间值不协调信息系统基于变精度优势关系的知识约简算法[J]. 曲阜师范大学学报(自然科学版),2018, 44(3):41-47.
WANG L. Knowledge reduction algorithm of interval-valued inconsistent information systems based on variable precision dominance relation[J]. Journal of Qufu Normal University(Natural Science), 2018, 44(3):41-47.

[5]　ZHANG X, MEI C L, CHEN D G, et al. Multi-confidence rule acquisition and confidence-preserved attribute reduction in interval-valued decision systems[J]. International Journal of Approximate Reasoning, 2014, 55(8):1787-1804.

[6]　唐鹏飞, 莫智文, 谢鑫. 区间值决策表中基于相对知识粒度的属性约简[J]. 重庆理工大学学报(自然科学), 2021, 35(11):286-292.
TANG P F, MO Z W, XIE X. Attribute reduction based on relative knowledge granularity in interval-valued decision tables[J]. Journal of Chongqing University of Technology(Natural Science), 2021, 35(11):286-292.

[7]　唐鹏飞, 张贤勇, 莫智文. 基于依赖度的区间集决策信息表属性约简[J]. 计算机应用研究, 2021, 38(11):3300-3303,3309.
TANG P F, ZHANG X Y, MO Z W. Attribute reduction of interval set decision information table based on

dependence degree [J]. Application Research of Computers, 2021, 38(11):3300-3303,3309.

[8] CHEN Y M, MIAO D Q, WANG R Z. A rough set approach to feature selection based on ant colony optimization[J]. Pattern Recognition Letters, 2010, 31(3):226-233.

[9] WANG J, MIAO D Q. Analysis on attribute reduction strategies of rough set[J]. Journal of Computer Science & Technology, 1998, 13(2):189-192.

[10] 刘正, 陈雪勤, 张书锋. 基于最小化邻域互信息的邻域熵属性约简算法[J]. 微电子学与计算机, 2020, 37(3):26-32.
LIU Z, CHEN X Q, ZHANG S F. Neighborhood entropy attribute reduction based on minimizing neghborhood mutual information [J]. Microelectronic & Computer, 2020, 37(3):26-32.

[11] LIU Y, XIE H, WANG L G, et al. Hyperspectral band selection based on a variable precision neighborhood rough set[J]. Applied Optics, 2016, 55(3): 462-472.

[12] HU Q H, ZHAO H, YU D R. Efficient symbolic and numerical attribute reduction with neighborhood rough sets[J]. Pattern Recognition & Artificial Intelligence, 2008, 21(6):732-738.

[13] 贾俊芳, 张英. 基于属性重要度的变精度邻域粗糙集知识约简[J]. 山西大同大学学报(自然科学版), 2014, 30(6):1-3,27.
JIA J F, ZHANG Y. Knowledge reduction of variable precision nerighborhood rough set based on attribute importance degree[J]. Journal of Shanxi Datong University(Natural Science Edition), 2014, 30(6):1-3, 27.

[14] 吴荣, 张文娟, 李进金. 对象导出三支概念格的熵属性约简[J]. 华侨大学学报(自然科学版), 2021, 42(5):694-700.
WU R,ZHANG W X,LI J J. Entropy attribute reduction of object-induced three-way concept lattice[J]. Journal of Huaqiao University (Natural Science), 2021,42(5):694-700.

[15] 李明, 甘秀娜, 王月波. 基于集成学习的决策粗糙集特定类属性约简算法[J]. 计算机应用与软件, 2021, 38(6):262-270.
LI M, GAN X N, WANG Y B. Class specific attribute reduction algorithm of decision rough set based on ensemble learning [J]. Computer Applications and Software, 2021, 38(6):262-270.

(编 辑 李 波)