基于样本类别的邻域粗糙集正域计算

彭潇然[1,2]，刘遵仁[2]，纪　俊[2]

PENG Xiaoran[1], LIU Zunren[2], JI Jun[2]

1. 青岛大学 数据科学与软件工程学院，山东 青岛 266071
2. 青岛大学 计算机科学技术学院，山东 青岛 266071
1. College of Data Science and Software Engineering, Qingdao University, Qingdao, Shandong 266071, China
2. College of Computer Science and Technology, Qingdao University, Qingdao, Shandong 266071, China

PENG Xiaoran, LIU Zunren, Ji Jun. A positive region computation of neighborhood rough set based on category of samples. Computer Engineering and Applications

Abstract：For an attribute reduction algorithm based on the neighborhood rough set, the positive region calculation is the necessary basis of its efficient performance and the uppermost part of its time cost. And the speed of the calculation is mainly determined by measure times between samples. In the condition of ensuring the correctness of the calculation, the less the measure times are, the faster the calculation is. In existing positive region calculations, there are usually a large measure times between samples that have the same category. Aimed at this case, this paper firstly proves that the measure between samples that have the same category is meaningless to the positive region calculation in neighborhood rough set. Then according to the proof, a positive region calculation based on category of samples is proposed. Compared with an existing positive region calculation, the experimental result shows that this proposed calculation is effective and faster. And this calculation is more suitable for data sets with fewer categories of samples.

Key words：rough set; neighborhood rough set; positive region computation; attribute reduction; category

摘　要：对基于邻域粗糙集的属性约简算法而言，正域计算是保证其有效性的重要依据，也是影响其时间开销的最主要部分。正域计算的速度主要由样本间度量计算的次数决定。在确保正确性的条件下，样本间度量计算的次数越少，则正域计算越快。在现有的正域计算中，通常存在着大量同类别样本间的度量计算。针对这个现象，首先证明在邻域粗糙集的正域计算中，同类别样本间的度量计算对正域计算是无贡献的，然后据此提出了基于样本类别的正域计算。和现有的正域计算相比，实验结果表明，该正域计算有效且更快速。而且，该正域计算更适用于样本类别数较少的数据集。

关键词：粗糙集；邻域粗糙集；正域计算；属性约简；样本类别
文献标志码：A　**中图分类号**：TP18　doi：10.3778/j.issn.1002-8331.1705-0347

基金项目：国家自然科学基金(No.61503208)。

作者简介：彭潇然(1994 -)，男，研究生，研究领域为粗糙集理论，E-mail: pxr1203@qq.com；刘遵仁(1963 -)，男，博士，副教授，研究领域为粗糙集理论、数据挖掘、智能计算等；纪俊(1982 -)，男，博士，副教授，研究领域为数据挖掘、大数据技术、转化医学等。

1 引言

粗糙集理论认为知识是有粒度的，是一种对论域中对象进行分类的能力。经典的 Pawlak 粗糙集[1]采用等价划分和等价类的概念保证了粒度计算的进行，但是这种处理方式只适用于离散型变量，而现实应用中需要处理的数据类型往往是数值型的，这种局限滞缓了粗糙集理论的应用。针对这个问题，Zadeh[2]提出了信息粒化和粒度计算的概念。Lin[3]在信息粒化、粒度的基础上提出了邻域模型的概念。Hu[4]提出了基于邻域粒化和粗糙逼近的决策表属性约简算法。经各方研究后提出的邻域粗糙模型可以处理数值型数据，这大大拓展了粗糙集理论的应用范围[2-7]。

但是，和 Pawlak 粗糙集的正域计算不同，因为引进了邻域粒化的概念，在邻域粗糙集的正域计算中，邻域信息粒子需要通过度量计算来确定，这导致邻域实数空间下的计算量要比经典离散空间下的计算量大得多。正域计算直接影响着基于邻域粗糙集算法的时间开销[8]。

在“信息爆炸”的时代，能有效且快速地对信息进行处理是十分有意义的。作为粗糙集理论的重要应用之一，属性约简一直以来是国内外学者研究的热点[9-18]。其中，基于邻域粗糙集，为了能快速地得到属性约简的结果，Hu[4]提出了基于前向贪心思想的属性约简算法(Fast forward heterogeneous attribute reduction based on neighborhood rough sets，F2HARNRS)。通过改进 F2HARNRS 算法的正域计算，Liu[8]随后提出了时间开销更少的快速属性约简算法 (Fast hash attribute reduction algorithm, FHARA)。

基于 Hu[4]、Liu[8]的研究，本文针对 F2HARNRS 算法和 FHARA 算法中正域计算的不足，提出了基于样本类别的正域计算，然后在此正域计算的基础上，进一步提出了基于样本类别的快速属性约简算法 (Fast Attribute Reduction Algorithm Based on Category，FARABC)。通过和 FHARA[8]算法进行比较，实验证明 FARABC 算法能有效且更快速地得到数据集的属性约简，在最好的情况下，算法的时间开销能缩减 5 倍左右。

2 相关概念[4]

2.1 邻域粒化

定义 1(度量计算) 给定 n 维实数空间 R^n，对于空间中的任意两个点 $x_i=(x_{i1},x_{i2}\ldots,x_{in})$ 和 $x_j=(x_{j1},x_{j2},\ldots,x_{jn})$，定义 $d(x_i,x_j)$ 是 R^n 上的一个度量计算，满足：

$$d(x_i,x_j)=\left(\sum_{p=1}^{n}\left|x_{ip}-x_{jp}\right|^2\right)^{\frac{1}{2}}$$

定义 2(邻域信息粒子) 在实数空间上，定义样本的非空有限集合 $U=\{x_1,x_2\ldots,x_n\}$，且称 U 为论域。定义 U 上的样本 x_i 的 δ - 邻域为 $\delta(x_i)=\{x_j\mid x_j\in U,d(x_i,x_j)\le\delta\}$，其中 $\delta\ge 0$。$\delta(x_i)$ 称作由 x_i 生成的 δ-邻域信息粒子，简称为 x_i 的邻域粒子。

2.2 邻域决策系统及上下近似

定义 3(邻域决策系统) 定义四元组 $\mathrm{NDT}=(U,C\cup D,V,f)$ 为邻域决策系统。其中 U 是论域；C 是条件属性集；D 是决策属性集，且 $C\cap D=\varnothing$，$C\neq\varnothing$，$\mathrm{D}\neq\varnothing$；$V$ 是信息函数 f 的值域。

定义 4(上下近似) 对于一个给定的邻域决策表 $\mathrm{NDT}=(U,C\cup D,V,f)$，$D$ 将 U 划分为 N 个等价类：$D_1,D_2,\ldots,D_N$，$\forall B\in C$，定义决策属性集 D 关于 B 的下近似和上近似为：

$$\underline{N_BD}=\bigcup_{i=1}^{N}\underline{N_BD_i}\ ,$$

$$\overline{N_BD}=\bigcup_{i=1}^{N}\overline{N_BD_i}$$

其中，

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

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

根据定义 1：

$\delta_B(x_i) = \{x \mid d(B(x_i), B(x)) \leq \delta, x \in U\}$

定义决策属性集 D 关于 B 的正域为 $Pos_B(D) = \underline{N_B D}$ ，边界域为 $BND_B(D) = \overline{N_B D} - \underline{N_B D}$ ，负域为 $NEG_B(D) = U - \overline{N_B D}$ 。

3 基于样本类别的正域计算

根据定义 4 可知，对于 n 维实数样本空间上的论域 U 和样本 $x_i \in U$ ，经典的正域计算包括对样本 x_i 的 δ-邻域计算和对 $\delta(x_i) \subseteq D_i$ 的判别，而基于样本类别的正域计算仅依赖样本 x_i 的 δ-邻域计算。其中，δ-邻域计算于依赖 x_i 与其它样本间的度量计算。度量计算越少，正域计算越快。

3.1 F2HARNRS 和 FHARA 算法的正域计算

在文献[4]中，Hu 给出了 F2HARNRS 算法。分析其正域计算：如图 1 所示，在对样本 x_i 的正域判定中，与 x_i 进行度量计算的待选样本是 U 中的所有样本。算法正域计算的时间复杂度为 $O(n|U|^2)$ 。这种正域邻域计算是经典的正域计算。

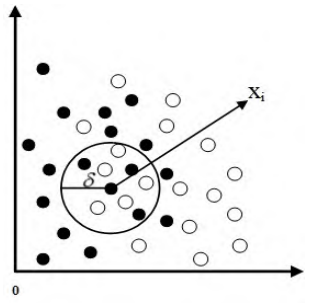

图 1 邻域实数空间下 F2HARNRS 算法中 x_i 的计算量

在文献[8]中，Liu 提出了比 F2HARNRS 算法更快速的 FHARA 算法。在其正域计算中，提出了一种映射划分策略：首先根据各样本与标准样本间的距离大小给各样本划分等级，然后基于等级将各样本映射到不同的有限集合 $B_0, B_1, ..., B_k$ 中，其中 $B_k = \{x_i \mid x_i \in U, k = \lceil d(x_i, x_0)/\delta \rceil\}$ ，x_0 是标准样本，定义 $x_0 = \{\forall a \in C, a(x_0) = \min[a(x_i)]\}$ ，$x_i \in U$ 。

分析其正域计算：基于图 1 中的样本分布，如图 2 所示，在对样本 x_i 的正域判定中，与 x_i 进行度量计算的待选样本是 x_i 己身所在扇环及相邻扇环中的样本。设样本均匀分布在样本空间中，则该正域计算的时间复杂度为 $O(q \cdot n|U|^2)$ ，其中 $q = 4/\lceil \max d(x_i, x_0)/\delta \rceil$ 。

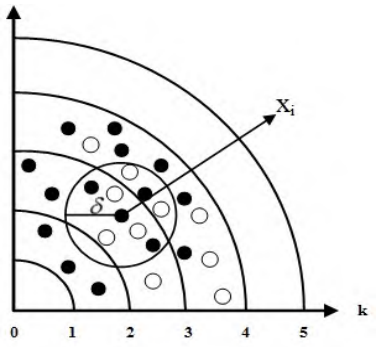

图 2 邻域实数空间下 FHARA 算法中 x_i 的计算量

相较 F2HARNRS 算法，FHARA 算法通过映射划分的策略缩减了样本间度量计算的次数，从而提高了正域计算的计算速度。

需注意到，因为同类别的样本在空间分布上往往相邻，所以在以上算法的正域计算中，存在相当多的同类别样本之间的度量计算。但是，根据本文的分析，这种度量计算对正域计算是无贡献的。

3.2 本文的正域计算

针对以上正域计算中存在的现状，提出定理 1。

定理 1：在邻域粗糙集的正域计算中，同类别样本之间的度量计算对正域计算是无贡献的。

证明：由定义 4 可知：对于样本 $x_i \in U$ ，若 $x_i \in \underline{N_B D_i}$ ，即 $x_i \in Pos_B(D)$ ，则其 δ-邻域 $\delta_B(x_i) \subseteq D_i$ 。

进一步做等价推导：若 $x_i \in Pos_B(D)$ ，则 $\forall x_j \in \delta_B(x_i)$ ，满足 $D(x_i) = D(x_j)$ ，即 $\delta_B(x_i)$ 中所有样本均是同一类别。

进一步做逆否推导：若 $\exists x_j \in \delta_B(x_i)$ ，满足 $D(x_i) \neq D(x_j)$ ，即 $\delta_B(x_i)$ 中存在某两个样本不是同一类别，则 $x_i \notin Pos_B(D)$ 。

由此得出结论：x_i 是否属于 $Pos_B(D)$ ，关键在于判断 $\delta_B(x_i)$ 中是否存在某个与 x_i 类别不同的样本，即

其与同类别样本之间的度量计算对正域计算是无贡献的。即证。

据此定理 1，改进 1。

改进 1:基于图 1 中的样本分布，如图 3 所示，在对样本 x_i 的正域判定中，与 x_i 进行度量计算的待选样本是 $x_j \in \{x \mid D(x) \neq D(x_i), x \in U\}$，若存在 x_j 使 $d(x_i, x_j) \leq \delta$ 成立，则立即判定 $x_i \notin Pos_B(D)$，否则 $x_i \in Pos_B(D)$。

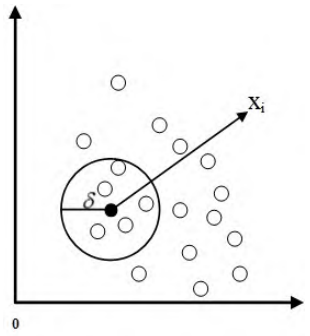

图 3　邻域实数空间下改进 1 中 x_i 的计算量

在改进 1 中，虽然通过样本类别筛除了同类样本，但在其异类样本中仍存在不少可避免计算的样本。采用以下策略作进一步筛除：对于 n 维的样本 x_i 和 x_j，称某 1 维上的计算为粗略计算，n 维上的计算为精细计算。显然，粗略计算比精细计算具有更少的时间开销，但其不作为正域判定的依据。在进行精细计算之前先作粗略计算，若不满足粗略计算的条件则直接跳过精细计算，否则再进行精细计算。度量计算 $d(x_i, x_j)$ 是精细计算。

定理 2：对于 n 维样本 $x_i = (x_{i1}, x_{i2} \ldots, x_{in}) \in U$ 和 $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn}) \in U$，若 $\exists |x_{ip} - x_{jp}| > \delta$，$p = 1,2,...,n$，则 $x_j \notin \delta(x_i)$。

证明：根据题设不妨设 $|x_{i1} - x_{j1}| > \delta$，即 $|x_{i1} - x_{j1}|^2 > \delta^2$，且 $|x_{ip} - x_{jp}|^2 \geq 0$，$p = 2,3,...,n$，则 $d(x_i, x_j) = \left(\sum_{p=1}^{n} |x_{ip} - x_{jp}|^2 \right)^{\frac{1}{2}} > \delta$，易知若 $d(x_i, x_j) > \delta$，则 $x_j \notin \delta(x_i)$。即证。

改进 2:在改进 1 的基础上，根据定理 2 提出粗略计算 $|x_{i1} - x_{j1}| \leq \delta$。基于图 1 中的样本分布，如图 4 所示，在对样本 x_i 的正域判定中，与 x_i 进行度量计算的待选样本是 $x_j \in \{x \mid D(x) \neq D(x_i), |x_{i1} - x_{x1}| \leq \delta, x \in U\}$，若存在 x_j 使 $d(x_i, x_j) \leq \delta$，则立即判定 $x_i \notin Pos_B(D)$，否则 $x_i \in Pos_B(D)$。

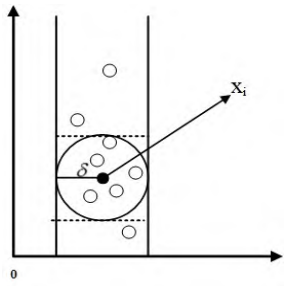

图 4　邻域实数空间下改进 2 中 x_i 的计算量

如图 4 虚线部分所示，理论上，若对样本 x_i 的每一维均做粗略计算，能进一步筛除需要做精细计算的样本。但需注意到，随着粗略计算次数的增多，粗略计算筛除样本的效果呈指数递减，其时间开销呈线性接近精细计算。粗略计算毕竟不作为正域判定的依据，若满足粗略计算的条件还需进行精细计算，过于追求筛选效果，反而会增大时间开销。第一次粗略计算的筛除效果是最好的，其时间开销却是最小的。所以我们仅做 1 维上的粗略计算，保证此策略有效性的最大化。

基于改进 2，分析 F2HARNRS 算法和 FHARA 算法：在对样本 x_i 的正域判定中，若 x_i 与 x_j 做度量计算，且满足 $d(x_i, x_j) \leq \delta$，此时判定 $x_i \notin Pos_B(D)$。在这种情况下，注意到，在对样本 x_j 的正域判定中，肯定也有 $d(x_j, x_i) \leq \delta$ 成立，即判定 $x_i \notin Pos_B(D)$ 时可同时判定 $x_j \notin Pos_B(D)$。

根据以上分析，提出本文的基于样本类别的正域计算。和经典的正域计算相比，这种正域计算不必判定 $\delta(x_i) \subseteq D_i$，仅依赖 δ-邻域计算。

定义 5(基于样本类别的正域计算)：对于一个给定的邻域决策系统 $\mathrm{NDT} = (U, C \cup D, V, f)$，根据样本类别，$D$ 将 U 划分为 N 个等价类：$D_1, D_2, ..., D_N$。$\forall B \subseteq C$，在对样本 $x_i \in D_i$ 进行正域判定时，若已判定 $x_i \notin Pos_B(D)$，则跳过计算，否则与待选样本

$x_j \in U - D_i$ 按照改进 2 的策略进行度量计算，若 $d(x_i, x_j) \le \delta$ ，则立即判定 $x_i \notin Pos_B(D)$ ，且同时判定 $x_j \notin Pos_B(D)$ 。

3.3 性能分析

设对于 n 维实数样本空间上的论域 U ，D 将 U 划分为 N 个等价类：$D_1, D_2, ..., D_N$ ，且 $|D_1| = |D_2| = ... = |D_N| = \frac{|U|}{N}$ ，U 中有一半的样本属于正域，则基于样本类别的正域计算在进行计算时，其计算量如图 5 阴影所示。

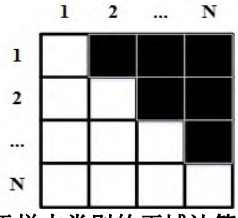

图 5 基于样本类别的正域计算的计算量

在图 5 中，阴影方格 $B(i, j)$ 的计算量是 $p \cdot n\left(\frac{|U|}{N}\right)^2$ ，其中 $p = 2\delta / \max(x_{p1} - x_{q1})$ ，$x_p, x_q \in D_j$ ，总共需计算 $\frac{N(N-1)}{2}$ 个方格，所以基于样本类别的正域计算的时间复杂度近似为 $O\left(\frac{P(N-1)}{2N} \cdot n|U|^2\right)$ ，其中 $p = 2\delta / \max(x_{p1} - x_{q1})$ ，$x_p, x_q \in U$ 。

在完成对数据集归一化处理的条件下，对于基于样本类别的正域计算，其时间复杂度的表达式中 $\max(x_{p1} - x_{q1}) \approx 1$ ，所以其时间复杂度可进一步表示为 $O\left(\frac{(N-1)\delta}{N} \cdot n|U|^2\right)$ ；对于 FHARA 算法的正域计算，其时间复杂度的表达式中 $\max d(x_i, x_0) \approx \sqrt{n}$ ，所以其时间复杂度可进一步表示为 $O(\frac{4\delta}{\sqrt{n}} \cdot n|U|^2)$ 。对比两个表达式可知，基于样本类别的正域计算除了受到 δ 的取值、样本维数和样本个数的影响外，还受到样本类别数的影响。

时间复杂度 $O\left(\frac{(N-1)\delta}{N} \cdot n|U|^2\right)$ 说明，若 δ 的取值和数据集的规模恒定，则基于样本类别的正域计算的时间开销会随着数据集中样本类别数 N 的增加而增大，当 N 很大时，上述表达式可表示为 $O\left(\delta \cdot n|U|^2\right)$ 。本文的实验部分也证明了这点：本文的 FARABC 算法对类别数较少的数据集进行属性约简的效率最高。

对于时间复杂度 $O(\frac{4\delta}{\sqrt{n}} \cdot n|U|^2)$ 需要注意的是，FHARA 算法的效率受样本在样本空间中分布情况的影响。上述表达式是在样本均匀分布的情况下求得的，而在实际中样本的分布往往边缘稀疏，中心密集，这种情况会限制了图 2 中扇环筛除样本的能力；其次，基于扇环映射的样本划分在进行正域判定前，需要对样本进行 hash 映射，在进行正域判定时，又需要进行 hash 访问，这些操作也存在一定的时间开销。

3.4 基于样本类别的正域计算算法

根据以上思路，设计基于样本类别的正域计算 $Pos(U, B, D, \delta)$ ，其中 $U = D_1 \cup D_2 \cup ... \cup D_N$ 。如算法 1 所示。

算法 1 ：

输入：U ，B ，D ，δ

输出：$\mathrm{Pos}_B(D)$

Step 1 初始化 Pos=U

Step 2 for each $x_i \in D_i, i = 1, 2, ..., N$

if $x_i \in \mathrm{Pos}$

for each $x_j \in U - D_i$

if $|x_{i1} - x_{j1}| \le \delta$

if $d(x_i, x_j) \le \delta$

$\mathrm{Pos} \leftarrow \mathrm{Pos} - \{x_i, x_j\}$;

break ;

end if

end if

end for

end if

end for

Step 3 return Pos

4 基于样本类别的快速属性约简算法

对一个数据集而言，如何删除冗余属性，得到

属性约简是粗糙集理论研究的重点之一。

定义 6[1](属性约简) :对于一个给定的邻域决策系统 $NDT=(U,C\cup D,V,f)$，$\forall B\subseteq C$，若满足 $\mathrm{Pos}_B(D)=\mathrm{Pos}_C(D)$，则称 B 是一个独立属性子集；如果对 $\forall a\in B$，$\mathrm{Pos}_{B-\{a\}}(D)<\mathrm{Pos}_B(D)$，则称 B 为 C 的一个属性约简。

贪心策略具有以较少时间求解最优解或次优解的特点。为了更快速地求得数据集的属性约简，F2HARNRS 算法[4]和 FHARA 算法[8]均采用了贪心策略：初始化属性约简集合为空集，当前正域为空集，每次选取使当前正域中样本个数增加最多的属性加入集合，直至对于当前集合而言所有属性的重要度全为 0 或样本全划入当前正域中时，输出集合。其中，根据新增加的属性不会使已属于正域的样本变为非正域样本这一性质，在算法的计算过程中，每次仅对还未判定为正域的样本进行正域计算。如算法 2 所示。

算法 2[4,8]：

输入：决策表 $(U,C\cup D,V,f)$

输出；属性子集 red

Step 1 初始化 red=∅，待检验样本 smp_chk=U

当前正域 max_pos=∅，最好属性 max_i=∅

Step 2 while smp_chk ≠ ∅

max_pos = ∅

for each $k_i\in(C-\mathrm{red})$

$\mathrm{Pos_i}=\mathrm{Pos}(\mathrm{smp\_chk},\mathrm{red}\cup \mathrm{k_i},\mathrm{D},\delta)$;

if $|\mathrm{max\_pos}|<|\mathrm{Pos_i}|$

max_pos=$\mathrm{Pos_i}$;

max_i=$\mathrm{k_i}$;

end if

end for

if max_pos ≠ ∅

red=red∪max_i ;

smp_chk=smp_chk-max_pos ;

else

break ;

end if

end while

Step 3 return red

可以看出，正域计算 $Pos(U,B,D,\delta)$ 影响着算法 2 的时间开销。相比 F2HARNRS 算法，FHARA 算法的工作在于改进了正域计算，从而减少了算法时间开销。本文沿用算法 2，结合本文的基于样本类别的正域计算(算法 1) ,提出基于样本类别的快速属性约简算法(Fast Attribute Reduction Algorithm Based on Category，FARABC)。

在该算法下，假设某一数据集有 m 个属性，约简结果中包含 k 个属性，且每增加一个属性正域中增加 $\frac{|U|}{k}$ 个样本，则进行正域判定的次数为：

$$\begin{aligned} & m|U|+(m-1)|U|\frac{k-1}{k}+\ldots+(m-k)|U|\frac{1}{k} \\ &<\frac{m|U|(1+2+\ldots+k)}{k} \\ &=\frac{m|U|(1+k)}{2} \end{aligned}$$

5 实验分析

5.1 实验环境及方案

UCI（University of California Irvine）提供了一系列用于测试的标准数据集。本文从 UCI 中选取了 12 个数据集，并按其样本数的大小进行升序排列，如表 1 所示。

表 1 数据集描述

编号	数据集	样本数	属性数	类别数
1	Zoo	101	16	7
2	Iris	150	4	3
3	Wine	178	13	3
4	Sonar	208	60	2
5	Ionosphere	351	34	2
6	Libras movement	360	90	15
7	WDBC	569	30	2
8	Credit Approval	690	14	2
9	German Credit	1000	19	2
10	CMC	1473	9	3

11	Segmentation	2310	19	7
12	Abalone	4177	7	28

本次实验在一台 Intel(R) Pentium(R) CPU G620 和 4GB 内存的 PC 机上，采用 Windows 7 环境下的 MATLAB R2016b 进行实验。

通过和 FHARA 算法进行对比，实验将从三个方面对 FARABC 算法进行分析。首先，通过比较两种算法得到的属性约简，分析 FARABC 算法的有效性；其次，通过比较两种算法的运行时间和样本间的度量计算次数，分析 FARABC 算法的时间开销；最后，通过比较 FARABC 算法相对于 FHARA 算法在度量计算次数上的比例和数据集中样本类别数的关系，分析 FARABC 算法的效率。其中，FHARA 算法和 FARABC 算法的比较实质上是两种正域计算的比较。

5.2 实验结果

5.2.1 FARABC 算法的有效性

为了去掉量纲对数据的影响，先对数据集数据进行归一化处理。参考 Hu[4]和 Liu[8]的实验结果，本文在区间 $(0,0.2]$ 中随机地选取 δ 的取值。将 FHARA 算法和 FARABC 算法各执行十次，取十次运行时间中的最小值作为算法的运行时间。实验运行结果如表 2 所示。

表 2 算法得到的属性约简

数据集	δ	FHARA	FARABC
Zoo	0.08	4,13,12,6,8	4,13,12,6,8
Iris	0.12	4,3,2,1	4,3,2,1
Wine	0.11	13,10,7,4,2	13,10,7,4,2
Sonar	0.19	58,1,45,37,17,13,10,19	58,1,45,37,17,13,10,19
Ionosphere	0.18	1,5,27,12,24,9,34,3,8,7,17	1,5,27,12,24,9,34,3,8,7,17
Libras movement	0.03	56,83,18,13	56,83,18,13
WDBC	0.1	23,8,22,25,29,19,10	23,8,22,25,29,19,10
Credit Approval	0.16	14,7,10,2,6,5,8,3,9,4,11,1,13,12	14,7,10,2,6,5,8,3,9,4,11,1,13,12
German Credit	0.01	2,10,4,1,3,8	2,10,4,1,3,8
CMC	0.16	4,7,8,3,2,1,6,5,9	4,7,8,3,2,1,6,5,9
Segmenta-tion	0.03	19,2,17,1,18,14	19,2,17,1,18,14
Abalone	0.08	3,6,5,7,4,2,1	3,6,5,7,4,2,1

由表 2 可知，在 δ 取值相同的情况下，两种算法得到的约简结果一致，这说明了 FARABC 算法是正确有效的。

5.2.2 FARABC 算法的时间开销

两种算法的运行时间如表 3 所示。

表 3 算法的运行时间 s

编号	数据集	FHARA	FARABC
1	Zoo	0.153486	0.13305
2	Iris	0.082747	0.055788
3	Wine	0.41165	0.219212
4	Sonar	5.31434	2.226284
5	Ionosphere	12.850363	2.570889
6	Libras movement	4.237643	2.914878
7	WDBC	6.633978	1.798142
8	Credit Approval	15.05088	2.366395
9	German Credit	13.532281	1.734165
10	CMC	19.582114	3.229287
11	Segmentation	24.922299	14.291288
12	Abalone	34.01808	17.390264

表 3 的折线图如图 6 所示。

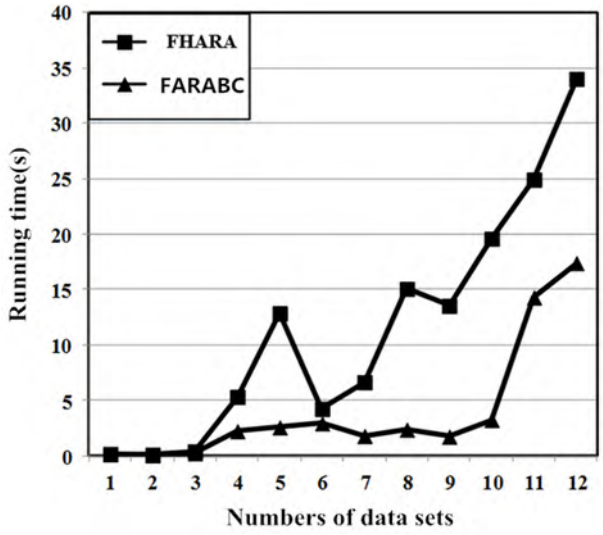

图 6　两种算法的运行时间

对比两种算法运行时间的折线，可以看出 FARABC 算法折线整体处于 FHARA 算法折线的下方，这表明 FARABC 算法比 FHARA 算法具有更少的时间开销。其次，相较 FARABC 算法，FHARA 算法折线波动大，这说明 FHARA 算法的效率受样本分布的影响，这点符合 3.3 节性能分析中的结论。

两种算法的度量计算的总次数如表 4 所示。

表 4　算法的度量计算次数　次

编号	数据集	FHARA	FARABC
1	Zoo	52674	43400
2	Iris	27469	5419
3	Wine	167660	48203
4	Sonar	2903815	920486
5	Ionosphere	8071229	1224745
6	Libras movement	1694566	854337
7	WDBC	3254243	675552
8	Credit Approval	8503765	1062440
9	German Credit	8104999	543487
10	CMC	10685460	1505227
11	Segmentation	8085310	5956201
12	Abalone	14293103	8546022

表 4 的柱状图如图 7 所示。

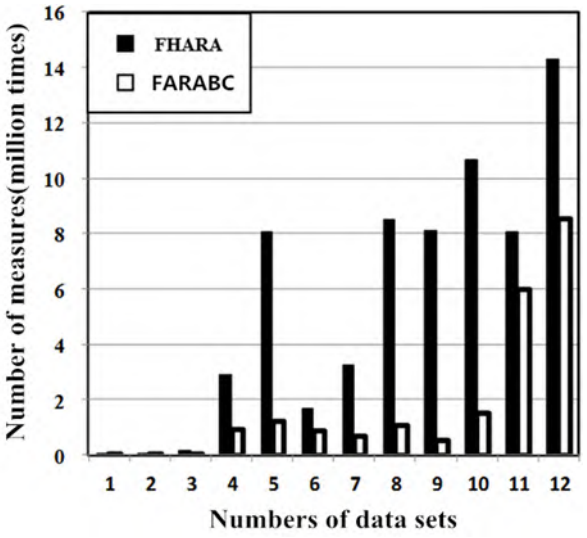

图 7　算法的度量计算次数

通过计算次数的对比，我们发现 FARABC 算法能明显地减少约简过程中度量计算的次数。

据上可知，和 FHARA 算法相比，FARABC 算法能有效且更快速地得到数据集的属性约简。

5.2.3　FARABC 算法的效率

由于运行时间受系统误差的影响，且度量计算次数直接影响着算法的时间开销，本部分的分析建立在两种度量计算的次数上。对于各数据集，将其在 FARABC 算法下得到的计算次数除以其在 FHARA 算法下得到的计算次数，用得到的比值表示 FARABC 算法相对于 FHARA 算法的效率，其比值越低，说明 FARABC 算法的效率越高。将数据集按照类别数的大小进行升序排列，如表 5 所示。

表 5　度量计算次数的比值(单位:%)

编号	数据集	类别数	比值
1	Sonar	2	31.70
2	Ionosphere	2	15.17
3	WDBC	2	20.76
4	Credit Approval	2	12.49
5	German Credit	2	6.71
6	Iris	3	19.73

7	Wine	3	28.75
8	CMC	3	14.09
9	Zoo	7	82.39
10	Segmentation	7	73.67
11	Libras movement	15	50.42
12	Abalone	28	59.79

表 5 折线图如图 8 所示。

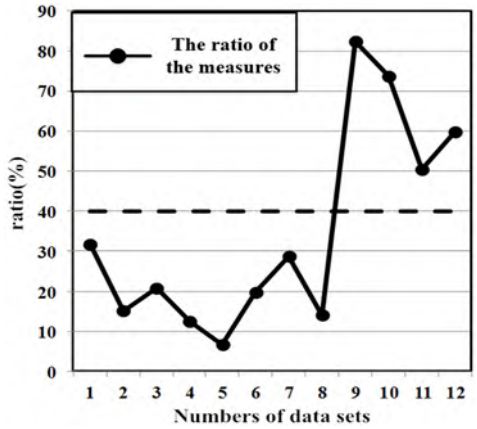

图 8　FARABC 算法的效率与 N 的关系

分析图 8 折线图可知，相对来说，前 8 个数据集的类别数 N 较小，其比值也较小，后 4 个数据集的类别数 N 较大，其比值也较大。

不失一般性地，通过以上分析可知：数据集的类别数 N 较小时，度量计算次数的比值较小，即 FARABC 算法相对 FHARA 算法的效率较高。这说明 FARABC 算法对类别数较少的数据集进行属性约简的效率最高，这点符合 3.3 节性能分析中的结论。

将前 8 个数据集得到的比值取平均值得 18.68，由此得出结论：相比 FHARA 算法的时间开销，FARABC 算法的时间开销最好能缩减 5 倍左右。

5.3　实验结论

上述结论即证明，基于样本类别的正域计算能有效且更快速地得到数据集的属性约简，且对类别数较少的数据集进行计算时效率最高。

6　结束语

本文提出了在邻域粗糙集的正域计算中，同类别样本间的度量计算对正域计算是无贡献的这一结论，进而提出了基于样本类别的正域计算。实验证明了该正域计算有效且更快速，但同时也分析了其适用的范围。对于基于邻域粗糙集的算法，特别是对于因迭代次数多、计算量大而造成时间开销大的算法，在处理样本类别数较少的数据集时都可以采用该正域计算进一步缩减算法的时间开销，优化算法的性能。针对类别数较多的数据集，如何进一步提高基于样本类别的正域计算的效率，我们将在后续的工作中对此问题进行研究。

参考文献:

[1] Pawlak Z, So-Winski R. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research, 1994, 72(3): 443- 459.

[2] Zadeh LA. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets and Systems, 1997, 90(90): 111-127.

[3] Lin TY. Granular Computing on binary relations I:Data mining and neighborhood systems[J]. Rough Sets in Knowledge Discovery, 1998, 18(1):107-121.

[4] Hu Q, Yu D, Liu J, Wu C. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences,2008,178(18):3577–3594.

[5] 王国胤. Rough 集理论与知识获取[M]．西安：西安交通大学出版社, 2001:147—156

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

[7] 胡清华,于达人.应用粗糙计算[M].北京:科学出版社, 2012.

[8] Liu Y, Huang W, Jiang Y, Zeng Z. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences, 2014, 271(7): 65-81.

[9] 刘勇,熊蓉,褚健. Hash 快速属性约简算法[J].计算机学报, 2009, 32(8): 1493-1499.

[10] Meng Z, Shi Z. A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets[J]. Information Sciences An International Journal, 2009, 179(16):2774-2793.

[11] 刘遵仁,吴耿锋.基于邻域粗糙模型的高维数据集快速约简算法[J].计算机科学,2012,39(10):268-271.

[12] 陈昊,杨俊安,庄镇泉.变精度粗糙集的属性核和最小属性约简算法[J].计算机学报,2012,35(5):1011-1017.

[13] Jia X, Liao W, Tang Z, Shang L. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2013, 219(1):151-167.

[14] Ye Dongyi, Chen Zhaojiong, Ma Shenglan. A novel and better fitness evaluation for rough set based minimum; attribute reduction problem[J].Information Sciences, 2013, 222(3):413-423.

[15] 丁卫平,王建东,管致锦.基于量子精英蛙的最小属性自适应合作型协同约简算法[J].计算机研究与发展, 2014, 51(4):743-753.

[16] Meng Z, Shi Z. On quick attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2016, 330(C):226-244.

[17] 潘瑞林,李园沁,张洪亮等.基于α信息熵的模糊粗糙属性约简方法[J].控制与决策,2017,32(2):340-348.

[18] 张云莉,范年柏.决策信息系统的增量式 F-并行属性约简[J].计算机工程与应用, 2017, 53(2):83-87.