基于邻域概率近似精度的属性约简

Attribute Reduction Based on Approximate Accuracy of Neighborhood Probability

周艳红

Yanhong Zhou

中国民用航空飞行学院理学院　中国·四川 广汉 618307

School of Science, Civil Aviation Flight Academy of China, Guanghan, Sichuan, 618307, China

摘　要：属性约简的目标是消除冗余和不相关属性，是优化处理和知识发现的基础，它在邻域概率粗糙集的研究中具有重要地位。论文首先利用邻域概率近似精度，提出属性约简及属性重要度的相关定义；其次，利用属性重要度构建分类属性约简的启发式算法；最后，通过实例对算法进行有效性说明。

Abstract: The goal of attribute reduction is to eliminate redundant and irrelevant attributes. It is the basis of optimization processing and knowledge discovery. It plays an important role in the research of neighborhood probabilistic rough sets. Firstly, based on the accuracy of neighborhood probability approximation, the related definitions of attribute reduction and attribute importance are proposed. Secondly, the heuristic algorithm of classification attribute reduction is constructed by using attribute importance. Finally, an example is given to illustrate the effectiveness of the algorithm.

关键词：邻域概率粗糙集；近似精度；分类属性约简；粒计算

Keywords: neighborhood probabilistic rough set; approximate accuracy; categorical attribute reduction; granular computing

基金项目：中国民用航空飞行学院面上项目（项目编号：J2019-037）。

DOI: 10.12346/sde.v4i3.6014

1 引言

粗糙集 [1] 是一种处理不确定性度量的有效数学工具，并应用于机器学习与人脸识别 [2] 等领域，对此很多学者对其进行研究，并取得大量成果 [3-4]。例如，文献 [3] 在概率粗糙集模型中提出基于期望粒度的三种单调不确定性度量，为属性约简奠定良好的基础；文献 [4] 利用双量化思想研究了基于精度和重要度的属性约简。

经典粗糙集是建立在等价关系与等价类的基础上，要求过于严格，在处理连续型数据方面具有局限性。而邻域粗糙集是粗糙集的一种推广，能够处理连续型数据，应用更为广泛，特别是邻域粗糙集的属性约简研究，取得了大量成果。例如，文献 [5] 用邻域知识粒度构造出一种邻域型信息系统的增量式属性约简算法；文献 [6] 基于邻域粗糙集，扩张构建邻域互补信息度量并研究其启发式属性约简；文献 [7] 提出一种邻域组合度量的启发式属性约简算法，是为了对混合型信息系统达到更好的属性约简性能；文献 [8] 引入邻域类的正确分类率，定义属性质量度，提出一种基于正域的增量和平均正确分类率的增率相结合的属性度量方法；文献 [9] 利用已有的信息度量，从层次视角研究基于决策表中观中层的特定类属性约简；文献 [10] 采用粒计算技术及相关的三层粒结构，构建具有粒化单调性的条件邻域熵，进而研究其相关属性约简。

在邻域粗糙集模型中，对噪声的容忍性相对较差，对一些实际应用具有一定的局限性。因此，文献 [11] 通过引入两个参数，构造出邻域概率粗糙集模型，并逐步提出邻域概率精度、粗糙度和近似精度三种具有单调性的度量，但没有进行属性约简相关研究，论文主要在此的基础上，利用具有粒化单调性的邻域概率近似精度在宏观高层建立属性约简算

【作者简介】周艳红（1982-），女，中国重庆人，博士，讲师，从事粗糙集与粒计算研究。

法。具体内容安排如下：第 1 节复习邻域概率粗糙集的基本概念；第 2 节，首先利用邻域概率近似精度在宏观高层给出属性约简和属性重要度的定义，其次利用属性重要度提出在宏观高层的属性约简的启发式算法，最后利用实例对属性约简算法进行验证。

2 预备知识

本小节复习邻域概率粗糙集的基本知识，包括邻域信息系统、邻域关系、邻域概率上下近似、邻域概率近似精度等[12]。

定义 1：$IS=(U,C,V,f,\delta)$ 称为邻域信息系统。其中，U 为非空有限论域；C 为非空属性集；V 为所有属性的值域，即 $V=\bigcup_{c\in C}V_c$（$V_c\in[0,1]$ 表示属性 c 所有取值的集合）；$\delta\in[0,1]$ 为邻域参数。特别地，$DS=(U,C\bigcup D,V,f,\delta)$ 表示邻域决策系统；其中，C 与 D 分别为条件属性集与决策属性集，$U/D=\{X_1, X_2\cdots,X_m\}$ 为 D 确定的决策等价分类.

论文主要涉及邻域决策系统 DS，针对 DS 与 IS，下设 $A,B\subseteq C$。

定义 2：在 IS 中，设 $x,y\in U$，$C=\{a_1,a_2,\cdots,a_n\}$，则 C 的距离函数为：

$$d_C(x,y)=\left(\sum_{i=1}^{n}\left|v(x,a_i)-v(y,a_i)\right|^p\right)^{\frac{1}{p}}$$

其中，$d_C(x,y)$ 为 Manhattan 距离若 $p=1$，论文主要采用 Manhattan 距离。

定义 3：在 IS 中，$x\in U$ 在 A 上的 δ 邻域为：$n_A^{\delta}(x)=\{y|x,y\in U,d_A(x,y)\le\delta\}$，$A$ 决定的邻域关系为：$NR_{\delta}(A)=\{(x,y)\in U\times U|d_A(x,y)\le\delta\}$。

定义 4：在 IS 中，$0\le\beta<\alpha\le1,A\subseteq C,X\subseteq U$，则 X 关于 A 的邻域概率下、上近似分别为：

$$\underline{apr}_{A\delta}^{(\alpha,\beta)}(X)=\left\{x\in U\middle|p\left(X/n_A^{\delta}(x)\right)\ge\alpha\right\}$$

$$\overline{apr}_{A\delta}^{(\alpha,\beta)}(X)=\left\{x\in U\middle|p\left(X/n_A^{\delta}(x)\right)>\beta\right\}$$

定义 5：在 DS 中，$0\le\beta<\alpha\le1,A\subseteq C,X\subseteq U,U/D=\{X_1,X_2,\cdots,X_m\}$，则 U/D 关于 A 的邻域概率近似精度为：

$$r_{A\delta}^{(\alpha,\beta)}(U/D)=\frac{\left|\underline{apr}_{A\delta}\left(\underline{apr}_{C\delta}^{(\alpha,\beta)}(U/D)\right)\right|}{\left|\overline{apr}_{A\delta}\left(\overline{apr}_{C\delta}^{(\alpha,\beta)}(U/D)\right)\right|}$$

定义 5 利用了邻域粗糙集和邻域概率粗糙集的上、下近似进行定义。

3 基于邻域概率近似精度的属性约简

本小节主要在邻域概率粗糙集中利用邻域概率近似精度，给出属性约简和属性重要度的相关定义，并构建了在宏观高层的基于属性重要度的属性约简算法。

定理 1：以下两个条件等价：

（s1）$r_{A\delta}^{(\alpha,\beta)}(U/D)\ne r_{(A-\{\alpha\})\delta}^{(\alpha,\beta)}(U/D),\forall a\in A$

（s2）$r_{A'\delta}^{(\alpha,\beta)}(U/D)\ne r_{A\delta}^{(\alpha,\beta)}(U/D),\forall A'\subset A$

证明：①$(s1\Rightarrow s2)$若 $\forall A'\subset A$，则 $\exists a\in A-A'\subset A$，s.t.，$A'\subseteq A-\{a\}\subset A$，由单调性知 $r_{A'\delta}^{(\alpha,\beta)}(U/D)\le r_{(A-\{\alpha\})\delta}^{(\alpha,\beta)}(U/D)<r_{A\delta}^{(\alpha,\beta)}(U/D)$。因此，$r_{A'\delta}^{(\alpha,\beta)}(U/D)\ne r_{A\delta}^{(\alpha,\beta)}(U/D)$，即 $(s1\Rightarrow s2)$ 成立。②$(s2\Rightarrow s1)$ $\forall a\in A$，设 $A'=A-\{a\}\subset A$，由 $s2$ 知，$s1$ 成立。得证。

定义 6：设 $A\subseteq C,$，且给定条件（s）：$r_{A\delta}^{(\alpha,\beta)}(U/D)=r_{C\delta}^{(\alpha,\beta)}(U/D)$。若 A 满足条件（s）与（s1）或条件（s）与（s2），则称为 C 的一个相对 U/D 约简，所有相对约简集记为 $\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)$。

定义 7：若 $r_{C\delta}^{(\alpha,\beta)}(U/D)\ne r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)$，则称 c 在 C 中是必要的，否则是不必要的。C 中所有必要属性组成的集合称为核，记为 $Core_{C\delta}^{(\alpha,\beta)}(U/D)$，即：

$$Core_{C\delta}^{(\alpha,\beta)}(U/D)=\left\{c\in C:r_{C\delta}^{(\alpha,\beta)}(U/D)\ne r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)\right\}$$

定理 2：$Core_{C\delta}^{(\alpha,\beta)}(U/D)=\bigcap_{A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)}A$。

证明：①若 $c\notin\bigcap_{A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)}A,$，则 $\exists A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)$，使得 $c\notin A,$，$r_{A\delta}^{(\alpha,\beta)}(U/D)=r_{C\delta}^{(\alpha,\beta)}(U/D)$，$A\subseteq C-\{c\}\subset C$，由粒化单调性知 $r_{C\delta}^{(\alpha,\beta)}(U/D)=r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)$，故 $c\notin Core_{C\delta}^{(\alpha,\beta)}(U/D)$，因此，$Core_{C\delta}^{(\alpha,\beta)}(U/D)\subseteq\bigcap_{A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)}A$。

②若 $c\notin Core_{C\delta}^{(\alpha,\beta)}(U/D)$，由定义 7 知，$r_{C\delta}^{(\alpha,\beta)}(U/D)=r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)$。$\exists A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)$，$A\subseteq C-\{c\}$，使得 $r_{A\delta}^{(\alpha,\beta)}(U/D)=r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)$，故 $r_{A\delta}^{(\alpha,\beta)}(U/D)=r_{C\delta}^{(\alpha,\beta)}(U/D)$，故 $A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)$。由 $A\subseteq C-\{c\}$ 知，$c\notin A$，从而 $c\notin\bigcap_{A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)}A$。因此，$\bigcap_{A\in\mathrm{Re}\,d_{C\delta}^{(\alpha,\beta)}(U/D)}A\subseteq Core_{C\delta}^{(\alpha,\beta)}(U/D)$。

综上①、②得证。

类似于经典情形，定义 7 给出了核的计算公式，定理 2 表明核在所有约简中，因此核可以作为构建属性约简的基础。鉴于属性重要度是属性约简的重要启发度量，下面首先利用邻域概率近似精度给出属性重要度的定义。

定义 8：设 $a\in C-A$，则属性 a 对于条件属性集 A 相对于 U/D 的重要度定义为：

$$Sig(a,A,D)=r_{(A\cup\{a\})\delta}^{(\alpha,\beta)}(U/D)-r_{A\delta}^{(\alpha,\beta)}(U/D)$$

属性重要度表明了在属性集 A 的基础上通过增加属性 $\{a\}$，在增加过程中邻域概率近似精度的变化情况。表 1 为一种基于属性重要度的属性约简的启发式算法，其中属性重要度是用于属性增加的启发式搜索。

表 1 启发式算法

算法 1 基于核的启发式属性约简算法
Input：决策表 $DS=(U,C\cup D,V,f,\delta)$，给定 $(\alpha,\beta),\delta$
Output：$A\in \mathrm{Re}d_{C\delta}^{(\alpha,\beta)}(U/D)$.
步骤：
Step 1 计算 $r_{C\delta}^{(\alpha,\beta)}(U/D)$, 设置 $Core_{C\delta}^{(\alpha,\beta)}(U/D)=\phi$.
Step 2 // 计算 $Core_{C\delta}^{(\alpha,\beta)}(U/D)$
Step 3 设 $Core_{C\delta}^{(\alpha,\beta)}(U/D)=\phi$
Step 4 for $\forall c\in C$ do
Step 5 计算 $r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)$
Step 6 if $r_{(C-\{c\})\delta}^{(\alpha,\beta)}(U/D)\neq r_{C\delta}^{(\alpha,\beta)}(U/D)$ then
Step 7 $Core_{C\delta}^{(\alpha,\beta)}(U/D)=Core_{C\delta}^{(\alpha,\beta)}(U/D)\cup\{c\}$.
Step 8 end if
Step 9 end for
Step 10 // Addition
Step 11 设 $A=Core_{C\delta}^{(\alpha,\beta)}(U/D), CA=C-A$
Step 12 While $r_{A\delta}^{(\alpha,\beta)}(U/D)\neq r_{C\delta}^{(\alpha,\beta)}(U/D)$, do
Step 13 for $\forall a_i\in CA$, do
Step 14 计算 $Sig(a_i,A,D)$
Step 15 end for
Step 16 if $Sig(a_0,A,D)=\arg\max Sig(a_i,A,D)$，then
Step 17 $A=A\cup\{a_0\}, CA=CA-\{a_0\}$，计算 $r_{A\delta}^{(\alpha,\beta)}(U/D)$
Step 18 end if
Step 19 end while
Step 20 Return A

该算法基于邻域概率近似精度进行了求核计算。具体地，通过逐个删除条件属性，计算邻域概率近似精度，通过收集不相等的属性即得 $Core_{C\delta}^{(\alpha,\beta)}(U/D)$。在核基础上，采用属性重要度 $Sig(a,A,D)$ 进行启发式搜索，快速增加条件属性来获得一个属性约简。具体地，步骤 11~18 循环增加属性直到邻域概率近似精度值保持，其中主要加入具有最大属性重要度的属性，以加速整个算法的收敛速度。

下面通过实例以有效说明邻域概率近似精度的启发式约简算法 .

例 1：（如表 2）设 $DS=(U,C\cup D,V,f,\delta)$，其中 $U=\{x_1,x_2,\cdots x_6\}$，$C=\{a_1,a_2,a_3,a_4\}$，

$U/D=\{X_1,X_2\}=\{\{x_1,x_4,x_6\},\{x_2,x_3,x_5\}\}, \alpha=0.4,\beta=0.3,\delta=0.4.$

表 2 例 1 决策表

U	a_1	a_2	a_3	a_4	D
x_1	1	1	0.7	0.5	N
x_2	0.8	0.9	0.6	0.5	Y
x_3	0.5	0.9	0.5	0.5	Y
x_4	0.3	0.7	0.2	0.2	N
x_5	0.8	0.6	0.8	0.8	Y
x_6	0.9	0.8	0.6	0.5	N

具体步骤如下：

①求 $Core_{C,0.4}^{(0.4,0.3)}(U/D)$。

$r_{C,0.4}^{(0.4,0.3)}(U/D)=\frac{1}{2}$，$r_{(C-a_1),0.4}^{(0.4,0.3)}(U/D)=\frac{2}{11}$，$r_{(C-a_2),0.4}^{(0.4,0.3)}(U/D)=\frac{1}{2}$，$r_{(C-a_3),0.4}^{(0.4,0.3)}(U/D)=\frac{1}{2}$，$r_{(C-a_4),0.4}^{(0.4,0.3)}(U/D)=\frac{1}{2}$。

因为 $r_{C,0.4}^{(0.4,0.3)}(U/D)\neq r_{(C-a_1),0.4}^{(0.4,0.3)}(U/D)$，$r_{C,0.4}^{(0.4,0.3)}(U/D)=r_{(C-a_2),0.4}^{(0.4,0.3)}(U/D)$，$r_{C,0.4}^{(0.4,0.3)}(U/D)=r_{(C-a_3),0.4}^{(0.4,0.3)}(U/D)$，$r_{C,0.4}^{(0.4,0.3)}(U/D)=r_{(C-a_4),0.4}^{(0.4,0.3)}(U/D)$。

所以 $Core_{C,0.4}^{(0.4,0.3)}(U/D)=\{a_1\}$。

②求属性约简集。

由①得到 $Core_{C,0.4}^{(0.4,0.3)}(U/D)=\{a_1\}$，并令其为 A，$r_{A,0.4}^{(0.4,0.3)}(U/D)=0.0000$，因为 $r_{C,0.4}^{(0.4,0.3)}(U/D)\neq r_{A,0.4}^{(0.4,0.3)}(U/D)$，计算 $r_{(A\cup\{a_i\}),0.4}^{(0.4,0.3)}(U/D)(i=2,3,4)$：$r_{(A\cup\{a_2\}),0.4}^{(0.4,0.3)}(U/D)=0.0833$，$r_{(A\cup\{a_3\}),0.4}^{(0.4,0.3)}(U/D)=0.0000$，$r_{(A\cup\{a_4\}),0.4}^{(0.4,0.3)}(U/D)=0.1818$，由此得到：

$$Sig(a_2,A,D)=r_{(A\cup\{a_2\}),0.4}^{(0.4,0.3)}(U/D)-r_{A,0.4}^{(0.4,0.3)}(U/D)=0.0833$$

$$Sig(a_3,A,D)=r_{(A\cup\{a_3\}),0.4}^{(0.4,0.3)}(U/D)-r_{A,0.4}^{(0.4,0.3)}(U/D)=0.0000$$

$$Sig(a_4,A,D)=r_{(A\cup\{a_4\}),0.4}^{(0.4,0.3)}(U/D)-r_{A,0.4}^{(0.4,0.3)}(U/D)=0.1818$$

由此可知，A 更新为 $\{a_1,a_4\}$。

$r_{A,0.4}^{(0.4,0.3)}(U/D)=0.1818$，故 $r_{C,0.4}^{(0.4,0.3)}(U/D)\neq r_{A,0.4}^{(0.4,0.3)}(U/D)$，$r_{(A\cup\{a_i\}),0.4}^{(0.4,0.3)}(U/D)(i=2,3)$：$r_{(A\cup\{a_2\}),0.4}^{(0.4,0.3)}(U/D)=0.5000$，$r_{(A\cup\{a_3\}),0.4}^{(0.4,0.3)}(U/D)=0.5000$。

从而有：

$$Sig(a_2,A,D)=r_{(A\cup\{a_2\}),0.4}^{(0.4,0.3)}(U/D)-r_{A,0.4}^{(0.4,0.3)}(U/D)=0.3182,$$

$$Sig(a_3,A,D)=r_{(A\cup\{a_3\}),0.4}^{(0.4,0.3)}(U/D)-r_{A,0.4}^{(0.4,0.3)}(U/D)=0.3182$$

故 A 更新为 $\{a_1,a_2,a_4\}$ 或 $\{a_1,a_3,a_4\}$。

因为

$r_{C,0.4}^{(0.4,0.3)}(U/D)=r_{\{a_1,a_2,a_4\},0.4}^{(0.4,0.3)}(U/D), r_{C,0.4}^{(0.4,0.3)}(U/D)=r_{\{a_1,a_3,a_4\},0.4}^{(0.4,0.3)}(U/D)$，

所以约简集为 $\mathrm{Re}d_{C,0.4}^{(0.4,0.3)}(U/D)=\{\{a_1,a_2,a_4\},\{a_1,a_3,a_4\}\}$。

4 结语

论文通过在邻域概率粗糙集中所构建的邻域概率近似精度基础上，提出宏观高层的基于属性重要度的属性约简算法，并通过实例对属性约简启发式属性约简算法进行有效性验证。

参考文献

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

[2] Liu D, Li T R, Li H X. A multiple-category classification approach

with decision-theoretic rough sets[J]. Funda- menta Informaticae, 2012,115(2):173-188.

[3] G Y Wang, X A Ma, H Yu. Monotonic uncertainty measures for attribute reduction in probabilistic rough set model[J]. International Journal Approxinate Reasoning, 2015,5(9):41-67.

[4] Zhang X Y, Miao D Q. Double-quantitative fusion of accuracy and importance: Systematic measure mining, benign integration construction hierarchical attribute reduction[J]. Knowledge Based Systems, 2016,9(1):219-240.

[5] 陈曦,刘晶.基于邻域关系的知识粒度增量式属性约简算法[J].微电子学与计算机,2020,37(10):1-6.

[6] 陈帅,张贤勇,唐玲玉,等.邻域互补信息度量及其启发式属性约简[J].数据采集与处理,2020,35(4):630-641.

[7] 盛魁,卞显福,董辉,等.基于邻域粗糙集组合度量的混合数据属性约简算法[J].计算机应用与软件,2020,37(2):234-239.

[8] 李冬,蒋瑜,鲍杨婉莹.基于属性质量度的变精度邻域粗糙集属性约简[J].四川师范大学学报(自然科学版),2020,43(4):560-568.

[9] 周艳红,张迪,张强.基于单调信息度量的特定类属性约简[J].内江师范学院学报,2019,34(12):35-39.

[10] 周艳红,张贤勇,莫智文.粒化单调的条件邻域熵及其相关属性约简[J].计算机研究与发展,2018,55(11):2395-2405.

[11] 周艳红,张迪,莫智文.邻域概率粗糙集的不确定性度量[J].四川师范大学学报(自然科学版),2021,44(1):136-142.

[12] Hu Q H, Yu D R, Xie Z X. Neighborhood classifiers[J]. Expert Systems With Applications, 2008,3(4):866-876.