计算机科学
COMPUTER SCIENCE

平均近似精度的性质和应用

张夏苇, 孔庆钊

引用本文

张夏苇, 孔庆钊. 平均近似精度的性质和应用[J]. 计算机科学, 2024, 51(11A): 240300108-5.

ZHANG Xiawei, KONG Qingzhao. Properties and Applications of Average Approximation Accuracy[J]. Computer Science, 2024, 51(11A): 240300108-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300108

平均近似精度的性质和应用

张夏苇[1]　孔庆钊[2]
1 厦门理工学院数学与统计学院　福建 厦门 361024
2 集美大学理学院　福建 厦门 361021

摘　要　平均近似精度是粗糙集理论中新近提出的一个重要概念。首先分析平均近似精度的数学结构,给出平均近似精度一种新的解释;然后重点讨论平均近似精度的若干重要性质,相比传统方法,其能更有效地刻画粗糙集模型知识表示的能力;最后,探讨平均近似精度在不完备信息表和特征选择两方面的应用。这些研究成果丰富了粗糙集理论的内容,扩展了粗糙集理论在实际问题中的应用。

关键词:粗糙集;近似精度;属性约简;不完备信息表
中图分类号　TP182

Properties and Applications of Average Approximation Accuracy

ZHANG Xiawei[1] and KONG Qingzhao[2]
1 School of Mathematics and Statistics,Xiamen University of Technology,Xiamen,Fujian 361024,China
2 College of Science,Jimei University,Xiamen,Fujian 361021,China

Abstract　Average approximation accuracy is an important concept in rough set theory,which has only been proposed in recent years. In this paper,the mathematical structure of average approximation accuracy is first analyzed,and another new explanation for average approximation accuracy is provided. Then,we focus on discussing several important properties of average approximation accuracy,and find that average approximation accuracy can characterize the knowledge representation ability of rough set models more effectively than traditional methods. Finally,the applications of average approximation accuracy in incomplete information tables and feature selection are discussed,respectively. These research achievements will enrich the content of rough set theory and expand its application in practical problems.

Keywords　Rough set,Approximation accuracy,Attribute reduction,Incomplete information table

由于数据收集、传输和存储成本大幅度降低以及信息交互网络的快速发展,数据规模不断膨胀,数据标签日益复杂。为了有效分析海量复杂数据,基于不同的学习任务,人们提出了各种数据挖掘的方法,如模糊集合理论、神经网络方法、商空间理论、三支决策等[1-6]。为了更好地描述带有模糊边界的数据集合,Pawlak[7]提出了粗糙集理论,运用上下近似两个精确集逼近或近似描述这种边界模糊的集合。粗糙集理论分析数据最大的优势之一在于无需任何先验知识,所有参数都可从信息表的样本集中获得。因此,粗糙集理论被广泛应用于知识表示与发现、不确定推理、粒计算和特征选择等许多领域[8-12]。

在粗糙集理论中,给定一个信息表,根据该信息表诱导出的粒结构(划分、覆盖等)可以构建相应的粗糙集模型。那么,如何度量粗糙集模型近似描述知识的能力呢?在粗糙集理论中,近似精度被定义为下近似的基数和上近似的基数的商,可被用来度量任意目标概念被近似描述的准确程度[7]。近似精度越高说明该集合被描述得越准确,反之亦然。因此,近似精度是衡量粗糙集模型描述一个给定目标概念准确程度一个非常重要的指标。但是,传统的近似精度是一个局部概念,它会随着目标概念的改变而变化,无法客观度量一个粗糙集模型自身知识表示的水平。为此,Kong 等把信息表中所有论域子集近似精度的平均值定义为平均近似精度,用它来衡量一个粗糙集模型近似描述知识的能力[13]。与传统近似精度不同,平均近似精度是一个全局的概念,它只与信息表和粗糙集模型有关,不会随着论域子集的改变而改变。因此,平均近似精度能客观准确地度量粗糙集模型表示知识的能力。

本文首先简要回顾粗糙集和平均近似精度等相关概念;然后研究平均近似精度的数学结构,重点讨论平均近似精度的一些重要性质;最后探讨平均近似精度在不完备信息表和特征选择等方面的应用。

1　预备知识

本章简要介绍与粗糙集和平均近似精度相关的一些重要概念和知识。

一个信息表通常可表示为四元序组[14]:

$$I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$$

其中,OB 是一个包含所有对象的非空有限集,称为论域;AT 为一个包含所有属性的非空有限属性集;$V=\bigcup\limits_{a\in AT}V_a$,$V_a$ 是属性 a 的所有属性值的集合;$I_a:OB\to V_a$ 是一个信息函数,$\forall x\in OB$,$\forall a\in AT$,对象 x 关于属性 a 的属性值用 $I_a(x)$ 表示。

基金项目:福建省自然科学基金(2020J01707)
This work was supported by the Natural Science Foundation of Fujian Province,China(2020J01707).
通讯作者:孔庆钊(kongqingzhao@163.com)

对于任意属性集$A\subseteq AT$，可以定义OB上的一个等价关系$E_A: xE_Ay\Leftrightarrow I_a(x)=I_a(y), \forall a\in A$。$\forall x\in OB$，$x$的等价类表示为：$[x]_A=\{y\in OB\mid xE_Ay\}$。$\forall A\subseteq AT$，由属性集$A$确定的$OB$上的划分表示为：$OB/E_A=\{[x]_A\mid x\in OB\}$。

定义 1[7] 在信息表$I=(OB, AT, \{V_a: a\in AT\}, \{I_a: a\in AT\})$中，$A\subseteq AT$是一个属性子集，$\forall X\subseteq OB$，我们称

$$\underline{apr}_A(X)=\{x\in OB\mid [x]_A\subseteq X\}$$

$$\overline{apr}_A(X)=\{x\in OB\mid [x]_A\cap X\neq\varnothing\}$$

分别为集合X关于属性子集A的下近似和上近似。

我们称

$$\alpha_A(X)=\frac{|\underline{apr}_A(X)|}{|\overline{apr}_A(X)|}$$

为集合X关于属性子集A的近似精度，其中$|\cdot|$表示集合的基数。

从近似精度的定义可以看出，近似精度随着集合X的变化而改变。由此可见，定义在一个信息表上的粗糙集模型近似描述所有知识的能力是无法通过近似精度来度量的。为此，平均近似精度被引入来解决该问题。

定义 2[13] 在信息表$I=(OB, AT, \{V_a: a\in AT\}, \{I_a: a\in AT\})$中，$A\subseteq AT$是一个属性子集，我们称

$$\alpha_A(2^{OB})=\frac{\sum_{X\in 2^{OB}}\alpha_A(X)}{|2^{OB}|}$$

为粗糙集关于属性子集A的平均近似精度。

由定义 2 可以看出，平均近似精度是论域OB中所有子集近似精度的平均值。它与某个子集的近似精度无关，不会随着子集的改变而变化。平均近似精度只与信息表和粗糙集模型有关，可以很好地度量粗糙集模型描述知识的能力。

例 1 信息表$I=(\{I_a: a\in AT\}, \{I_a: a\in AT\})$，其中，$OB=\{x_1, x_2, x_3, x_4\}$，$OB/E_A=\{\{x_1, x_2\}, \{x_3, x_4\}\}$。论域$OB$有 16 个子集，分别为$X_1=\varnothing$，$X_2=\{x_1\}$，$X_3=\{x_2\}$，$X_4=\{x_3\}$，$X_5=\{x_4\}$，$X_6=\{x_1, x_2\}$，$X_7=\{x_1, x_3\}$，$X_8=\{x_1, x_4\}$，$X_9=\{x_2, x_3\}$，$X_{10}=\{x_2, x_4\}$，$X_{11}=\{x_3, x_4\}$，$X_{12}=\{x_1, x_2, x_3\}$，$X_{13}=\{x_1, x_2, x_4\}$，$X_{14}=\{x_1, x_3, x_4\}\}X_{15}=\{x_2, x_3, x_4\}$，$X_{16}=OB$。则

$$\alpha_A(2^{OB})=\frac{\sum_{i=1}^{16}\alpha_A(X_i)}{16}=0.375$$

定义 3 在信息表$I=(OB, AT, \{V_a: a\in AT\}, \{I_a: a\in AT\})$中，$A\subseteq AT$是一个属性子集，$\forall a\in A$，我们称

$$ID_a=1-\frac{\alpha_{A\setminus\{a\}}(2^{OB})}{\alpha_A(2^{OB})}$$

为属性a相对于属性集A的重要度，简称为属性a的重要度。

由定义 3 可以看出，属性a的重要度为属性a删除前后平均近似精度的改变量与属性a删除前平均近似精度的比值。属性a删除后引起的平均近似精度改变越大(小)，属性a的重要度就越大(小)，说明属性a越重要(不重要)。而且，$100\cdot ID_a\%$表示删除属性a后信息表(粗糙集模型)对所有知识表示所降低的能力占原信息表(原粗糙集模型)对所有知识表示能力的百分比。

2 平均近似精度的数学结构

本章研究平均近似精度的数学结构，并给出平均近似精度的另一种解释。借助近似精度，我们首先定义2^{OB}上的一个等价关系，进而得到2^{OB}上的一个划分。

在信息表$I=(OB, AT, \{V_a: a\in AT\}, \{I_a: a\in AT\})$中，$A\subseteq AT$是一个属性子集，$X, Y\in 2^{OB}$是两个论域子集，可定义论域幂集$2^{OB}$上的一个等价关系$E_{\alpha_A}$：

$$XE_{\alpha_A}Y\Leftrightarrow\alpha_A(X)=\alpha_A(Y)$$

$\forall X\in 2^{OB}$，集合X的等价类为：

$$[X]_{\alpha_A}=\{Y\in 2^{OB}\mid XE_{\alpha_A}Y\}$$

可得2^{OB}上的一个划分如下：

$$2^{OB}/E_{\alpha_A}=\{[X]_{\alpha_A}\mid X\in 2^{OB}\}=\{B_{1\alpha_A}, B_{2\alpha_A}, \cdots, B_{s\alpha_A}\}$$

(1) $\forall X_i\in B_{i\alpha_A}$，$\forall X_j\in B_{j\alpha_A}$，有

$\alpha_A(X_i)\neq\alpha_A(X_j)$，$i, j=1, 2, \cdots, s; i\neq j$。

根据式(1)，可得关于划分$2^{OB}/E_{\alpha_A}$的概率分布：

$$P_{\alpha_A}=\left(\frac{|B_{1\alpha_A}|}{|2^{OB}|}, \frac{|B_{2\alpha_A}|}{|2^{OB}|}, \cdots, \frac{|B_{s\alpha_A}|}{|2^{OB}|}\right)=(p_{1\alpha_A}, p_{2\alpha_A}, \cdots, p_{s\alpha_A}) \tag{2}$$

根据式(2)，平均近似精度可以被改写为：

$$\begin{aligned}\alpha_A(2^{OB})&=\frac{\sum_{X\in 2^{OB}}\alpha_A(X)}{|2^{OB}|}\\&=\sum_{i=1}^{s}\frac{|B_{i\alpha_A}|\cdot\alpha_A(X_i)}{|2^{OB}|}\\&=\sum_{i=1}^{s}\frac{|B_{i\alpha_A}|}{|2^{OB}|}\cdot\alpha_A(X_i)\\&=\sum_{i=1}^{s}p_{i\alpha_A}\cdot\alpha_A(X_i)\end{aligned} \tag{3}$$

其中，$X_i\in B_{i\alpha_A}$，$i=1, 2, \cdots, s$。

由式(3)，平均近似精度可以看作近似精度$\alpha_A(X_1)$，$\alpha_A(X_2)$，$\cdots$，$\alpha_A(X_s)$的加权平均数，权数分别是$p_{1\alpha_A}$，$p_{2\alpha_A}$，$\cdots$，$p_{s\alpha_A}$。

例 2 (接例 1)基于式(1)可得$2^{OB}/E_{\alpha_A}=\{B_{1\alpha_A}, B_{2\alpha_A}, B_{3\alpha_A}\}$，其中$B_{1\alpha_A}=\{\varnothing, \{x_1, x_2\}, \{x_3, x_4\}, OB\}$；$B_{2\alpha_A}=\{\{x_1, x_2, x_3\}, \{x_1, x_2, x_4\}, \{x_1, x_3, x_4\}, \{x_2, x_3, x_4\}\}$；$B_{3\alpha_A}=\{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_1, x_3\}, \{x_1, x_4\}, \{x_2, x_3\}, \{x_2, x_4\}\}$。而且$\forall X_1\in B_{1\alpha_A}$，有$\alpha_A(X_1)=1$；$\forall X_2\in B_{2\alpha_A}$，有$\alpha_A(X_2)=\frac{1}{2}$；$\forall X_3\in B_{3\alpha_A}$，有$\alpha_A(X_3)=0$。

根据式(1)，可得如下概率分布：

$$P_{\alpha_A}=(p_{1\alpha_A}, p_{2\alpha_A}, p_{3\alpha_A})=\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$$

根据式(2)，平均近似精度为

$$\begin{aligned}\alpha_A(2^{OB})&=\sum_{i=1}^{3}p_{i\alpha_A}\cdot\alpha_A(X_i)\\&=\frac{1}{4}\times 1+\frac{1}{4}\times\frac{1}{2}+\frac{1}{2}\times 0\\&=0.375\end{aligned}$$

由例 1 和例 2，可知平均近似精度可通过定义和数学期望两种方法表示或计算。

3 平均近似精度的性质

本章主要研究平均近似精度的若干重要性质。

定理 1 在信息表$I=(OB, AT, \{V_a: a\in AT\}, \{I_a: a\in AT\})$中，$A\subseteq AT$是一个属性子集，则$\frac{1}{2^{|OB|-1}}\leqslant\alpha_A(2^{OB})\leqslant 1$。

证明：当$OB/E_A=\{OB\}$时，可得：(1)$\alpha_A(\varnothing)=\alpha_A(OB)=1$；(2)对$\varnothing\subset X\subset OB$，有$\alpha_A(X)=0$。此时，平均近似精度取最小值，最小值为

$$\alpha_A(2^{OB})=\frac{2}{2^{|OB|}}=\frac{1}{2^{|OB|-1}}$$

当 OB/E_A 每个等价类都是单点集时,可得:$\forall X\subseteq OB$,有 $\alpha_A(X)=1$。此时,平均近似精度取最大值,最大值为 $\alpha_A(2^{OB})=\frac{2^{|OB|}}{2^{|OB|}}=1$。从而,可知 $\frac{1}{2^{|OB|-1}}\leqslant\alpha_A(2^{OB})\leqslant 1$。

定义 4[14] 对论域 OB 上两个划分 OB/E_1 和 OB/E_2,若 OB/E_1 中任何一个等价类都包含在 OB/E_2 中的某个等价类中,则称划分 OB/E_1 比划分 OB/E_2 细,记为 $OB/E_1\subseteq OB/E_2$。若 $OB/E_1\subseteq OB/E_2$ 且 $OB/E_1\neq OB/E_2$,则称划分 OB/E_1 比划分 OB/E_2 严格地细,记为 $OB/E_1\subset OB/E_2$。

从定理 1 中不难看出 $\alpha_A(2^{OB})$ 取最小值和最大值分别对应论域上最粗的划分(即不划分,原始对象集 OB 是一个等价类)和最细的划分(即每个对象形成的单点集是一个等价类)。而且基于定义 4,还有如下结论。

定理 2 在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A_1,A_2\subseteq AT$ 是两个属性子集,若 $OB/E_{A_1}\subseteq OB/E_{A_2}$,则

$$\alpha_{A_2}(2^{OB})\leqslant\alpha_{A_1}(2^{OB})$$

证明:如果 $OB/E_{A_1}=OB/E_{A_2}$,则 $\alpha_{A_1}(2^{OB})=\alpha_{A_2}(2^{OB})$。

如果 $OB/E_{A_1}\subset OB/E_{A_2}$,由定义 4,$\forall B_{11}\in OB/E_{A_1}$,存在 $B_2\in OB/E_{A_2}$ 使得 $B_{11}\subseteq B_2$。

若 $B_{11}\subset B_2$,$\forall x_1\in B_2/B_{11}$,存在 $B_{12}\in OB/E_{A_1}$ 使得 $x_1\in B_{12}$,则有 $B_{12}\subset B_2$。

若 $B_{11}\cup B_{12}\subset B_2$,$\forall x_2\in B_2/(B_{11}\cup B_{12})$,存在 $B_{13}\in OB/E_{A_1}$ 使得 $x_2\in B_{13}$,则有 $B_{13}\subset B_2$。

以上步骤一直持续下去,直到第 k 步,满足:$B_2=B_{11}\cup B_{12}\cup\cdots\cup B_{1k}$。由定义 1,可得 $\alpha_{A_2}(B_{1i})\leqslant\alpha_{A_1}(B_{1i})$,$i=1,2,\cdots,k$。

综上,可得

$$\alpha_{A_2}(2^{OB})\leqslant\alpha_{A_1}(2^{OB})$$

在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A\subseteq AT$ 是一个属性子集,且 a 是 A 中的一个属性。易知 $OB/E_{A/\{a\}}\subseteq OB/E_A$ 成立。根据定理 2,可得如下结论。

定理 3 在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A\subseteq AT$ 是一个属性子集,a 是 A 中的一个属性,则 $\alpha_{A/\{a\}}(2^{OB})\leqslant\alpha_A(2^{OB})$。

除了比较不同划分的粗细之外,还可以从其他角度对划分展开研究,比如 Wierman 根据划分的同构这一重要概念首次对一个论域上的所有划分进行讨论,定义了论域上所有划分的一个等价关系[15]。

定义 5[15] 对于论域 OB 上的两个划分 OB/E_1 和 OB/E_2,若存在一个双射 $f:OB/E_1\to OB/E_2$ 使得对 $\forall B\in OB/E_1$ 有 $|f(B)|=|B|$,则称划分 OB/E_1 和 OB/E_2 同构。

由划分同构的定义,不难得到如下结论。

定理 4 在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A_1,A_2\subseteq AT$ 是两个属性子集,若两个划分 OB/E_{A_1} 和 OB/E_{A_2} 同构,则 $\alpha_{A_1}(2^{OB})=\alpha_{A_2}(2^{OB})$。

例 3 在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A_1,A_2\subseteq AT$ 是两个属性子集,若 $OB=\{x_1,x_2,x_3,x_4\}$;则

$$OB/E_{A_1}=\{\{x_1,x_2\},\{x_3\},\{x_4\}\}$$

$$OB/E_{A_2}=\{\{x_1,x_3\},\{x_2\},\{x_4\}\}$$

由定义 5 可知,划分 OB/E_{A_1} 和 OB/E_{A_2} 同构,因此 $\alpha_{A_1}(2^{OB})=\alpha_{A_2}(2^{OB})=0.625$。

属性约简是粗糙集理论研究的核心内容之一。为了完成不同的学习任务,许多种类的属性约简被引入并被广泛研究[16-20]。保持粗糙集模型知识表示能力不变的属性约简是非常重要和常见的一种。

定义 6[16] 在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A\subseteq AT$ 是一个属性子集,若 A 满足条件:

(1)$OB/E_A=OB/E_{AT}$;

(2)$\forall a\in A$,有 $OB/E_{A\setminus\{a\}}\neq OB/E_{AT}$。

则称属性子集 A 是属性集 AT 的一个属性约简,记作 $reduct(AT)$,即 $reduct(AT)=A$。若 $a\in AT\text{-}reduct(AT)$,则称 a 为可约属性。

由定义 6 可知,可约属性被删除前后论域上的划分是相同的,即粗糙集模型的知识表示能力不变。而粗糙集模型的知识表示能力又可以用平均近似精度进行度量。于是,我们有如下重要结论。

定理 5 在信息表 $I=(OB,AT,\{V_a:a\in AT\},\{I_a:a\in AT\})$ 中,$A\subseteq AT$ 是一个属性子集,若 A 是属性集 AT 的一个属性约简当且仅当

(1)$\alpha_A(2^{OB})=\alpha_{AT}(2^{OB})$;

(2)$\forall a\in A$,有 $\alpha_{A\setminus\{a\}}(2^{OB})\neq\alpha_{AT}(2^{OB})$。

我们知道,属性约简侧重对粗糙集模型的知识表示能力进行定性分析,而平均近似精度着重从定量角度对粗糙集模型的知识表示能力进行描述。因此,定理 5 给出了属性约简和平均近似精度之间的一个等价刻画,将属性约简和平均近似精度完美地统一起来。

例 4 在表 1 所示的信息表中,$OB=\{x_1,x_2,\cdots,x_8\}$,$AT=\{a_1,a_2,a_3,a_4\}$。

表 1 信息表

Table 1 Information table

OB	a_1	a_2	a_3	a_4
x_1	1	1	1	1
x_2	1	1	1	1
x_3	1	1	0	0
x_4	1	0	0	1
x_5	1	1	0	0
x_6	0	1	0	0
x_7	0	1	0	0
x_8	0	0	0	1

由定义 6 可知,$A=\{a_1,a_2,a_3\}$ 是属性集 AT 的一个属性约简,即属性 a_4 被删除后论域上的划分没有改变。因此,粗糙集模型对知识的表示能力不变。而从平均近似精度的角度看,属性 a_4 被删除前后信息表的平均近似精度都是 0.4582,因此,粗糙集模型知识表示的能力是没有降低的。由此可见,属性约简和平均近似精度都可以判断粗糙集模型知识表示的能力是否发生变化。

但是,与传统方法相比,平均近似精度却能更有效地度量粗糙集模型知识表示的能力。

例 5(接例 4) 在信息表 1 中,令 $A_1=\{a_1,a_2\}$,$A_2=\{a_1,a_3\}$,有 $OB/E_{A_1}=\{\{x_1,x_2,x_3,x_5\},\{x_4\},\{x_6,x_7\},\{x_8\}\}$ $OB/E_{A_2}=\{\{x_1,x_2\},\{x_3,x_4,x_5\},\{x_6,x_7,x_8\}\}$。那么对划分 OB/E_{A_1} 和 OB/E_{A_2} 而言,基于哪个划分的粗糙集模型知识表示的能力更高一些呢?传统方法无法回答这个问题。

但是,由定义 2 可得,$\alpha_{A_1}(2^{OB})=0.2822$,$\alpha_2(2^{OB})=0.1851$。

由此可见，基于 OB/E_{A_1} 的粗糙集模型比基于 OB/E_{A_2} 的粗糙集模型有更高的知识表示能力。

4 平均近似精度的应用

平均近似精度是刻画粗糙集模型描述知识能力的一个有效度量，具有广泛的应用。本章尝试从不完备信息表中的数据填补和特征选择两个方面简要介绍平均近似精度的相关应用。

4.1 平均近似精度在不完备信息表中的应用

不完备信息表是数据挖掘中常见的一种数据表示系统。在不完备信息表中，有些数据的属性值是缺失的。处理不完备信息表中缺失的属性值最常用的一种方法就是基于某种规则对缺失的数据进行估计和填补，把一个不完备的信息表转变成一个完备的信息表[21-23]。在文献[24]中，Zhou 等介绍了估计缺失属性值的有效规则，如下所示：

(1)若对象 $x\in OB$ 在属性 $a\in AT$ 下的属性值缺失，则用该属性 a 下已知频率最高的属性值代替；

(2)若频率最高的属性值不只一个，则考虑其他属性对应的相同属性值的多少。选择相同属性值多的对象对应的属性 a 的属性值代替缺失的属性值。

此处，我们考虑在上述规则下进一步运用平均近似精度估计不完备信息表中缺失的数据。

案例 1 表 2 是一个带有 3 个缺失属性值的不完备信息表。如何合理估计表 2 中 3 个缺失的属性值 $*_1$，$*_2$，$*_3$ 呢？

表 2 一个不完备信息表

Table 2 Incomplete information table

OB	a_1	a_2	a_3
x_1	1	0	0
x_2	1	1	1
x_3	$*_1$	1	1
x_4	$*_2$	0	0
x_5	0	0	0
x_6	1	$*_3$	1

根据文献[24]中给出的规则，这 3 个缺失的属性值可以被合理地估计为：$*_1=0$ 或 1；$*_2=0$ 或 1；$*_3=0$。将这些估计的属性值填补到原信息表后可以得到 4 个完备的信息表。但是，基于这些信息表建立的粗糙集模型的知识表示的能力却是不尽相同的。下面，逐一进行分析：

情形 1 若 $*_1=0$，$*_2=0$，$*_3=0$，则 $OB/E_{AT}=\{\{x_1\},\{x_2,x_3\},\{x_4,x_5\},\{x_6\}\}$，可得 $\alpha_{AT}(2^{OB})=0.3625$。

情形 2 若 $*_1=0$，$*_2=1$，$*_3=0$，则 $OB/E_{AT}=\{\{x_1,x_4\},\{x_2,x_3\},\{x_5\},\{x_6\}\}$，可得 $\alpha_{AT}(2^{OB})=0.3625$。

情形 3 若 $*_1=1$，$*_2=0$，$*_3=0$，则 $OB/E_{AT}=\{\{x_1\},\{x_2\},\{x_3\},\{x_4,x_5\},\{x_6\}\}$，可得 $\alpha_{AT}(2^{OB})=0.4625$。

情形 4 若 $*_1=1$，$*_2=1$，$*_3=0$，则 $OB/E_{AT}=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\},\{x_6\}\}$，可得 $\alpha_{AT}(2^{OB})=0.4625$。

由此可知情形 3 和情形 4 不仅估计合理，而且对应的粗糙集模型知识表示的能力最强。此时，$*_1=1$，$*_2=0$，$*_3=0$ 或者 $*_1=1$，$*_2=1$，$*_3=0$ 是所有合理估计数据中最优的选择。

4.2 平均近似精度在特征选择中的应用

在粗糙集理论中，特征选择被称为属性约简，通过选取有代表性的或者重要的数据标签，优化数据信息系统，降低数据维度。特征选择能显著节省数据存储空间，有效降低时间消耗，有助于减少过拟合现象和避免维度灾难[25-30]。为了进行特征选择，需要判断各个特征(或属性)的重要程度，根据具体任务删除不重要或次要的特征。如何从知识表示的准确程度方面定量地描述属性的重要程度呢？平均近似精度是解决该问题的一个很好的工具。下面根据定义 3 中给出的属性重要度的概念，通过具体案例介绍平均近似精度在特征选择方面的一个具体应用。

案例 2(接例 4) 求信息表 1 中各个属性的重要度。

由表 1，关于属性集 AT 的划分为 $OB/E_{AT}=\{\{x_1,x_2\},\{x_3,x_5\},\{x_4\},\{x_6,x_7\},\{x_8\}\}$，则有 $\alpha_{AT}(2^{OB})=0.4582$；关于属性子集 $\{a_1,a_2,a_3\}$ 的划分为 $OB/E_{\{a_1,a_2,a_3\}}=\{\{x_1,x_2\},\{x_3,x_5\},\{x_4\},\{x_6,x_7\},\{x_8\}\}$，则有 $\alpha_{\{a_1,a_2,a_3\}}(2^{OB})=0.4582$；关于属性子集 $\{a_1,a_2,a_4\}$ 的划分为 $OB/E_{\{a_1,a_2,a_4\}}=\{\{x_1,x_2,x_3,x_5\},\{x_4\},\{x_6,x_7\},\{x_8\}\}$，则有 $\alpha_{\{a_1,a_2,a_4\}}(2^{OB})=0.2822$；关于属性子集 $\{a_1,a_3,a_4\}$ 的划分为 $OB/E_{\{a_1,a_3,a_4\}}=\{\{x_1,x_2\},\{x_3,x_4,x_5\},\{x_6,x_7,x_8\}\}$，则有 $\alpha_{\{a_1,a_3,a_4\}}(2^{OB})=0.1851$；关于属性子集 $\{a_2,a_3,a_4\}$ 的划分为 $OB/E_{\{a_2,a_3,a_4\}}=\{\{x_1,x_2\},\{x_3,x_5,x_6,x_7\},\{x_4,x_8\}\}$，则有 $\alpha_{\{a_2,a_3,a_4\}}(2^{OB})=0.1843$。

于是，$ID_{a_1}=0.5978$，$ID_{a_2}=0.5960$，$ID_{a_3}=0.3841$，$ID_{a_4}=0$。

因此，从知识表示准确度的角度来看，属性重要程度从大到小的属性依次为 a_1,a_2,a_3,a_4。从该案例可以看出，由平均近似精度定义的属性重要程度可以定量地准确描述各个属性在知识表示方面的重要程度，为做属性约简时属性的取舍提供了可靠的参考依据。

结束语 在粗糙集理论中，平均近似精度能有效度量粗糙集模型的知识表示能力。本文首先讨论了平均近似精度的数学结构，平均近似精度可被看作传统近似精度的加权平均数。然后探讨了平均近似精度许多重要性质，加深了对平均近似精度的理解；最后给出了平均近似精度在不完备信息表和特征选择两方面的应用。

同时，注意到平均近似精度的计算复杂度较高，设计有效算法以便提高计算平均近似精度的效率是今后值得研究的一个重要课题。基于平均近似精度的粗糙集模型和属性约简等也需要不断地研究和探索。

参 考 文 献

[1] ZADEH L. Fuzzy sets [J]. Information and Control, 1965, 8: 338-353.

[2] RAMESH DHANASEELAN F, JEYA SUTHA M. Detection of breast cancer based on fuzzy frequent itemsets mining [J]. IRBM, 2021, 42(3): 198-206.

[3] LEE H, HSIEH C J, LEE J S. Local critic training for model-parallel learning of deep neural networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4424-4436.

[4] ZHANG B, ZHANG L. Theory and applications of problem solving [M]. North Holland Publishing, Amsterdam, 1992.

[5] YAO Y Y. Three-way decisions with probabilisticrough sets [J]. Information Sciences, 2010, 180(3): 341-353.

[6] YAO Y Y. Three-way decision and granular computing [J]. In-

ternational Journal of Approximate Reasoning, 2018, 103: 107-123.

[7] PAWLAK Z. Rough sets [J]. International Journal of Computing and Information Science, 1982, 11(5): 341-356.

[8] KONG Q Z, ZHANG X W, XU W H, et al. A novel granular computing model based on three-way decision [J]. International Journal of Approximate Reasoning, 2022, 144: 92-112.

[9] KONG Q Z, ZHANG X W, XU W H, et al. Attribute reducts of multi-granulation information system [J]. Artificial Intelligence Review, 2020, 53(2): 1353-1371.

[10] XU W H, HUANG M, JIANG Z Y, et al. Graph-based unsupervised feature selection for interval-valued information system [J/OL]. https://doi.org/10.1109/TNNLS.2023.3263684.

[11] KONG Q Z, CHANG X E. Rough set model based on variable universe [J]. CAAI Transactions on Intelligence Technology, 2022, 7(3): 503-511.

[12] KONG Q Z, XU W H, ZHANG D X. A comparative study of different granular structures induced from the information systems [J]. Soft Computing, 2022, 26(1): 105-122.

[13] KONG Q Z, CHANG X E. Two kinds of average approximation accuracy [J]. CAAI Transactions on Intelligence Technology, 2024, 9(2): 481-490.

[14] PAWLAK Z. Information systems theoretical foundations [J]. Information Systems, 1981, 6(3): 205-218.

[15] WIERMAN M J. Measuring uncertainty in rough set theory [J]. International Journal of General Systems, 1999, 28(4/5): 283-297.

[16] PAWLAK Z. Rough sets. Theoretical aspects of reasoning about data [M]. Kluwer Academic Publishers, Dordrecht, 1991.

[17] MI J S, WU W Z, ZHANG W X. Approaches to knowledge reduction based on variable precision rough set model [J]. Information Sciences, 2004, 159(3/4): 255-272.

[18] MIAO D Q, ZHAO Y, YAO Y Y, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model [J]. Information Sciences, 2009, 179(24): 4140-4150.

[19] WANG F, LIANG J Y, QIAN Y H. Attribute reduction: a dimension incremental strategy [J]. Knowledge-Based Systems, 2013, 39: 95-108.

[20] YANG Y Y, CHEN D G, DONG Z. Novel algorithms of attribute reduction with variable precision rough set model [J]. Neurocomputing, 2014, 139: 336-344.

[21] YUAN J L, CHEN M, JIANG T, et al. Complete tolerance relation based parallel filling forincomplete energy big data [J]. Knowledge-Based Systems, 2017, 132: 215-225.

[22] CHEN J, SHAO J. Jackknife variance estimation for nearest-neighbor imputation [J]. Journal of the American Statistical Association. 2001, 96(453): 260-269.

[23] SALAMA A S, El-BARBARY O G. Topological approach to retrieve missing values in incomplete information systems [J]. Journal of the Egyptian Mathematical Society, 2017, 25: 419-423.

[24] ZHOU X Z, HUANG B, LI H X, et al. Rough set theory and method for knowledge acquisition in incomplete information systems[D]. Nanjing: Nanjing University Press, 2010.

[25] HU X, ZHANG H, YANG C M, et al. Regularized spectral clustering with entropy perturbation [J]. IEEE Transactions on Big Data, 2021, 7(6): 967-972.

[26] KONG Q Z, WANG W T, XU W H, et al. A method of data analysis based on division-mining-fusion strategy [J]. Information Sciences, 2024, 666: 120450.

[27] JENSEN R, QIANG S. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1457-1471.

[28] YAO Y Y. Interpreting concept learning in cognitive informatics and granular computing [J]. IEEE Transactions on Systems Man and Cybernetics B, 2009, 39(4): 855-866.

[29] ZHANG J B, LI T R, RUAN D, et al. A parallel method for computing rough set approximations [J]. Information Sciences, 2012, 194: 209-223.

[30] XU W H, YUAN K H, LI W T, et al. An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution [J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(1): 76-88.

ZHANG Xiawei, born in 1981, master, associate professor. Her main research interests include granular computing, artificial intelligence and network diagnostic.

KONG Qingzhao, born in 1978, Ph.D, associate professor. His main research interests include granular computing and artificial intelligence.