文章编号：2095-2163(2021)10-0020-06 中图分类号：TP18 文献标志码：A

基于近似条件熵的集值决策表属性约简算法

唐鹏飞[1,2]

（1 四川师范大学 数学科学学院，成都 610066；2 四川师范大学 智能信息与量子信息研究所，成都 610066）

摘 要：集值决策表拓展了经典决策表，但其现有属性约简算法中属性重要度度量方式单一。针对集值决策表，采用近似条件熵提出属性约简及其启发式约简算法。将近似精度与条件信息熵进行信息融合，定义近似条件熵，证明粒化单调性等性质；提出基于近似条件熵的属性约简，设计启发式约简算法；采用集值决策表实例与数据实验进行有效验证。实验结果表明：与现有算法相比，提出算法不仅能够得到更优的约简结果，而且具有更高的分类精度。

关键词：粗糙集；集值决策表；近似条件熵；属性约简；启发式约简算法

Attribute reduction algorithm for set-valued decision tables based on approximation conditional entropy

TANG Pengfei[1,2]

（1 School of Mathematical Sciences, Sichuan Normal University, Chengdu 610066, China;
2 Institute of Intelligent Information and Quantum Information, Sichuan Normal University, Chengdu 610066, China）

【Abstract】 Set-valued decision tables extend classic decision tables, but the attribute significance measurement is single in their attribute reduction algorithms. Aiming at set-valued decision tables, this paper proposes attribute reduction and its heuristic algorithm by adopting approximation condition entropy. By fusing approximate accuracy and conditional information entropy, it defines the approximate conditional entropy. Based on the new information measure, it proposes the attribute reduction, and accordingly designeds a heuristic algorithm. Finally, the relevant validity is verified by set-valued decision table examples and data sets experiments. The experimental results show that the proposed attribute reduction algorithm can obtain smaller reduction results and higher classification accuracy than the current algorithms.

【Key words】 rough set; set-valued decision tables; approximation condition entropy; attribute reduction; heuristic reduction algorithm

0 引 言

属性约简是粗糙集理论的核心内容之一，主要是在保持相同分类能力的前提下进行冗余属性删除，从而达到数据表的优化处理[1]。近似条件熵作为一种由近似精度与条件信息熵信息融合得到的强健度量模型，已广泛应用于不确定性测量、属性约简、机器学习。对于经典决策表，文献[2]首先提出基于近似条件熵的属性约简；进一步，文献[3-4]将近似条件熵推广到邻域粗糙集；文献[5]利用近似条件熵进行电务设备故障预测；文献[6]分解近似条件熵，提出邻域粗糙集的特定类属性约简。可见，近似条件熵是信息表进行属性约简的有效手段，具有重要的现实应用意义。

集值决策表是经典决策表的一种扩展，近年来已经开展了一系列相关研究。例如，文献[7]采用条件信息熵构建集值系统属性约简。文献[8]基于相容关系，提出信息熵、粗糙熵、知识粒度来刻画集值信息系统的不确定性。文献[9]采用信息量构建集值信息系统属性约简。文献[10]基于集值信息系统上的优势关系，提出多粒度优势关系粗糙集模型。文献[11]基于相容关系，研究了一致集值决策信息系统中的属性约简和判断问题。文献[12]针对集值数据提出 2 种模糊粗糙近似，并分别构建其相对正域约简。文献[13]从知识粒度视角，提出集值信息系统的动态属性约简算法。文献[14]基于巴氏距离提出 λ 容差关系，研究了概率集值信息系统的动态变精度粗糙集模型。文献[15]针对集值决策信息系统中数据动态变化，构建增量式属性约简算法。文献[16]采用正域构建集值信息系统快速约简算法。文献[17]基于模糊相容关系，引入依赖度构建集值信息系统属性约简。归纳可见，现有

作者简介：唐鹏飞(1996-)，男，硕士研究生，主要研究方向：粗糙集、粒计算。

通讯作者：唐鹏飞 Email：2959194437@qq.com

收稿日期：2021-08-02

的集值决策信息系统属性约简仅从代数角度或信息角度单方面出发。前者刻画属性对论域中确定分类子集的影响，后者刻画属性对论域中不确定分类子集的影响[18]，存在局限性。

针对上述问题，本文引入近似条件熵，构建集值决策表的属性约简。具体地，基于相容关系，将近似精度与条件信息熵信息融合建立强健的近似条件熵度量模型，挖掘度量的粒化单调性，从而自然构建基于近似条件熵的属性约简及启发算法，最后提供实例分析与实验验证。

1　集值决策表的相关知识

集值决策表 $SVDT=\{U,AT=C\cup D,V,f\}$。其中，$U=\{x_1,x_2,\cdots,x_n\}$ 为非空有限论域；C 为条件属性集，D 为决策属性集，$C\cap D=\varnothing$；属性值域 $V=\cup_{a\in AT}V_a$，这里 V_a 为任意属性 $a\in AT$ 的值域；信息函数 $f:U\times AT\to V_a$，并赋予 x 在属性 a 上的值 $f(x,a)\in V_a$。对任意条件属性子集 $B\subseteq C$，定义 U 上相容关系：$T_B=\{(x,y)\in U\times U\mid f(x,a)\cap f(y,a)\neq\varnothing,\forall a\in B\}$，在 T_B 下，对象 $x\in U$ 关于 B 的相容类为 $TC_B(x)=\{y\in U\mid (x,y)\in T_B\}$，所有相容类构成一个覆盖 $U/T_B=\{TC_B(x)\mid x\in U\}$。决策属性集 D 在 U 上的划分 $U/T_D=\{D_1,\cdots,D_m\}$，而且 $T_D=\{(x,y)\in U\times U\mid f(x,D)=f(y,D)\}$，$D_h(1\leqslant h\leqslant m)$ 表示决策类。本文用 $|\cdot|$ 表示集合的基数。

定义 1[7]　决策划分 U/T_D 关于 B 的条件信息熵为：

$$HE(D\mid B)=-\sum_{i=1}^{|U|}p(TC_B(x_i))\sum_{h=1}^{m}p(D_h\mid TC_B(x_i))\times\log_2 p(D_h\mid TC_B(x_i)) \tag{1}$$

其中，$p(TC_B(x_i))=\dfrac{|TC_B(x_i)|}{|U|}$，$p(D_h\mid TC_B(x_i))=\dfrac{|TC_B(x_i)\cap D_h|}{|TC_B(x_i)|}$。

定义 2[16]　决策类 D_h 关于 B 的下、上近似集分别为：

$$\underline{BD}_h=\{x\in U\mid TC_B(x)\subseteq D_h\}, \tag{2}$$

$$\overline{BD}_h=\{x\in U\mid TC_B(x)\cap D_h\neq\varnothing\}, \tag{3}$$

决策类 D_h 关于 B 的近似精度为：

$$\alpha_B(D_h)=\frac{|\underline{BD}_h|}{|\overline{BD}_h|} \tag{4}$$

定理 1[7,16]决策分析原理可表示为：

(1) $B\subseteq C\Rightarrow\alpha_B(D_h)\leqslant\alpha_C(D_h)$；

(2) $B\subseteq C\Rightarrow HE(D\mid C)\leqslant HE(D\mid B)$

定义 1 提供条件信息熵，其仅能刻画粒化结构的不确定性；定义 2 提供近似精度，其仅能度量近似分类的不确定性[19]。这 2 种单一度量模型都存在一定的局限性。因此，本文将两者进行信息融合，构建一种更为全面的度量模型，设计出一种更优的约简算法。

2　基于近似条件熵的启发式属性约简算法

本节在条件信息熵与近似精度基础上，构建基于近似条件熵的属性约简及其启发式约简算法，并提供实例说明算法的有效性。为此，首先通过信息融合提出近似条件熵。

定义 3　决策划分 U/T_D 关于 B 的近似条件熵为：

$$AHE(D\mid B)=-\sum_{h=1}^{m}\log_2(2-\alpha_B(D_h))\times\sum_{i=1}^{|U|}p(TC_B(x_i))\times p(D_h\mid TC_B(x_i))\times\log_2 p(D_h\mid TC_B(x_i)) \tag{5}$$

近似条件熵引入近似精度作为条件信息熵的权重系数，有效融合了两者的优点，变得更为强健。既能度量近似分类的不确定性，又能表征粒化结构的不确定性，是一种更加全面的度量模型。

定理 2　$B\subseteq C\Rightarrow AHE(D\mid C)\leqslant AHE(D\mid B)$

证明　因为 $B\subseteq C$，由定理 1 可得 $\alpha_B(D_h)\leqslant\alpha_C(D_h)$，$HE(D\mid C)\leqslant HE(D\mid B)$。结合定义 3 可得 $AHE(D\mid C)\leqslant AHE(D\mid B)$。证毕。

推论 1　$AHE(D\mid B)\in[0,|U|\log_2|U|]$。特别地，当 U/T_B 最细时（即 $\forall x\in U,TC_B(x)=\{x\}$），$AHE(D\mid B)$ 取得最小值为 0。当 U/T_B 最粗时（即 $\forall x\in U,TC_B(x)=U$），且决策划分最细时（即 $\forall D_h\in U/T_D,|D_h|=1$），$AHE(D\mid B)$ 取得最大值为 $|U|\log_2|U|$。

定理 2 表明，通过信息融合得到的近似条件熵仍具有粒化单调性。进而，推论 1 给出相应的值域与最值条件。接下来，给出属性的必要性和独立性定义。

定义 4　$\forall a\in C,AHE(D\mid C)\neq AHE(D\mid C-\{a\})$，则称 a 为 C 中 D 必要的属性，否则称 a 为 C 中 D 不必要的属性。

定义 5　若 $\forall a\in C$ 为 C 中 D 必要的属性，则称 C 为 D 独立的。由 C 中所有 D 必要的属性组成的集合，称为 C 中 D 的核，记为 $Core_C(D)$。

基于上述近似条件熵的度量语义与粒化单调性,给出以下近似条件熵约简的定义。

定义 6 $B \subseteq C$ 为基于近似条件熵的属性约简,若能满足 2 个条件:

(1) $AHE(D \mid B) = AHE(D \mid C)$;

(2) $\forall a \in B, AHE(D \mid B - \{a\}) \neq AHE(D \mid B)$

这里,属性约简模拟经典情况,主要依托近似条件熵这一核心度量及其粒化单调性。定义 6 中的 2 条分别对应"联合充分性"和"个体必要性"。特别地,近似条件熵的粒化单调性可以挖掘启发式信息;由此,下面提出对应的属性重要度,并建立启发式约简算法。

定义 7 $B \subseteq C$ 且 $a \in B$, 则 a 对于 B 关于决策划分 U/T_D 的内属性重要度为:

$$SIG_{inner}(a,B,D) = AHE(D \mid B - \{a\}) - AHE(D \mid B) \tag{6}$$

$B \subseteq C$, $\forall a \in C - B$, 则 a 对于 B 关于决策划分 U/T_D 的外属性重要度为:

$$SIG_{outer}(a,B,D) = AHE(D \mid B) - AHE(D \mid B \cup \{a\}) \tag{7}$$

内属性重要度 $SIG_{inner}(a,B,D)$ 刻画在 B 中删除属性 a 之后近似条件熵值的增加量,而外属性重要度 $SIG_{outer}(a,B,D)$ 刻画在 B 上添加属性 a 之后近似条件熵值的减少量。相关度量值变化越大,则说明该属性越重要,因此两者提供了快速约简的属性选择机制。根据核属性概念(定义 5),下面利用内属性重要度构造一个求核方法。

定理 3 $\forall a \in C$, 则

a 为 C 中 D 必要的属性 $\Leftrightarrow SIG_{inner}(a,C,D) > 0$, 从而,

$$Core_C(D) = \{a \in C \mid SIG_{inner}(a,C,D) > 0\} \tag{8}$$

证明 若 a 为 C 中 D 必要的属性,则 $AHE(D \mid C) \neq AHE(D \mid C - \{a\})$, 由近似条件熵的粒化单调性知, $AHE(D \mid C) < AHE(D \mid C - \{a\})$, 因此 $SIG_{inner}(a,C,D) > 0$。反之,若 $SIG_{inner}(a,C,D) > 0$, 则 $AHE(D \mid C - \{a\}) > AHE(D \mid C)$, 从而 $AHE(D \mid C) \neq AHE(D \mid C - \{a\})$。由定义 4 知, a 为 C 中 D 必要的属性,故式(8)成立。证毕。

依据上述近似条件熵及约简的定义,下面以定义 7 的 2 种属性重要度为启发式信息,开发一个以核为约简起点的启发式约简算法,从而快速获取一个属性约简。算法步骤具体如下。

算法 1 基于近似条件熵的启发式约简算法

输入:集值决策表 $SVDT = \{U, AT = C \cup D, V, f\}$

输出:该区间集决策信息表的一个约简 R

步骤 1 计算 $AHE(D \mid C)$。

步骤 2 设置 $Core_C(D) = \varnothing$, $\forall a \in C$, 计算内属性重要度 $SIG_{inner}(a,C,D)$, 若 $SIG_{inner}(a,C,D) > 0$, 则 $Core_C(D) = Core_C(D) \cup \{a\}$, 得到 C 中 D 的核 $Core_C(D)$, 令 $R = Core_C(D)$。

步骤 3 计算决策划分 U/D 关于 R 的近似条件熵。若 $AHE(D \mid R) = AHE(D \mid C)$, 则执行步骤 5, 否则执行步骤 4。

步骤 4 $\forall a \in (C - R)$, 计算外属性重要度 $SIG_{outer}(a,R,D)$, 靠前选择外属性重要度最大的条件属性 a^* 并入 R 中,即进行更新 $R \leftarrow R \cup \{a^*\}$。如果此时有 $AHE(D \mid R) > AHE(D \mid C)$, 则重复该步骤选择与更新过程,直到达到条件 $AHE(D \mid R) = AHE(D \mid C)$, 才执行步骤 5。

步骤 5 向前遍历 R 中每个属性 a, 若 $AHE(D \mid R - \{a\}) = AHE(D \mid R)$, 则设置 $R \leftarrow R - \{a\}$。

步骤 6 返回 R。

算法 1 是以核为约简起点的启发式约简算法。步骤 2 通过内属性重要度寻找到核属性集,步骤 3 对其进行评估,若步骤 2 得到的子集 R 的近似条件熵大于全集 C 的(即 $AHE(D \mid R) > AHE(D \mid C)$),则需进入步骤 4 通过外属性重要度对剩余属性进行启发式搜索,并通过顺序选取最优属性以快速完成添加。可见,临近步骤 5 的 R 必然满足约简"联合充分性",但不一定满足约简"个体必要性"。由此,步骤 5 采取后项删除过程,以确保获得"个体必要性",从而 R 是一个基于近似条件熵的属性约简,最终被有效输出。

3 实例分析

本节通过一个实例,计算与分析近似条件熵性质及对应属性约简。

例 1 集值决策表见表 1。其中, $U/T_D = \{D_1, D_2\} = \{\{x_1,x_2,x_3,x_4\}, \{x_5,x_6,x_7,x_8\}\}$。

表 1 集值决策表

Tab. 1 Set-valued decision table

U	a_1	a_2	a_3	a_4	a_5	D
x_1	{1,2}	{1}	{1}	{2}	{2}	0
x_2	{1}	{1}	{1}	{2}	{1}	0
x_3	{2}	{1,2}	{2}	{1,2}	{1}	0
x_4	{1,2}	{1}	{2}	{1,3}	{2}	0
x_5	{1,2}	{1,2}	{2}	{3}	{2}	1
x_6	{1,2}	{2}	{2}	{1}	{1}	1
x_7	{1}	{1,2}	{1,2}	{1,2}	{1}	1
x_8	{1}	{2}	{1,2}	{1,3}	{1}	1

基于表 1,构造自然属性增链：

$B_1=\{a_1\}\subset B_2=\{a_1,a_2\}\subset B_3=\{a_1,a_2,a_3\}\subset B_4=\{a_1,a_2,a_3,a_4\}\subset C=\{a_1,a_2,a_3,a_4,a_5\}$。

作为例子,选取属性增链中的链元 B_2。计算其诱导的相容类为：

$TC_{B_2}(x_1)=\{x_1,x_2,x_3,x_4,x_5,x_7\}$,

$TC_{B_2}(x_2)=\{x_1,x_2,x_4,x_5,x_7\}$,

$TC_{B_2}(x_3)=\{x_1,x_3,x_4,x_5,x_6\}$,

$TC_{B_2}(x_4)=\{x_1,x_2,x_3,x_4,x_5,x_7\}$,

$TC_{B_2}(x_5)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$,

$TC_{B_2}(x_6)=\{x_3,x_5,x_6,x_7,x_8\}$,

$TC_{B_2}(x_7)=\{x_1,x_2,x_4,x_5,x_6,x_7,x_8\}$,

$TC_{B_2}(x_8)=\{x_5,x_6,x_7,x_8\}$,

由此,可得决策类 D_1、D_2 关于 B_2 的近似精度分别为：

$$\alpha_{B_2}(D_1)=\frac{|B_2D_1|}{|\overline{B_2}D_1|}=\frac{0}{7},\alpha_{B_2}(D_1)=\frac{|B_2D_1|}{|\overline{B_2}D_1|}=\frac{1}{8},$$

下面分别计算决策类 D_1、D_2 关于 B_2 的条件信息熵为：

$$HE(D_1\mid B_2)=-[\frac{4}{8}\log_2\frac{4}{6}+\frac{3}{8}\log_2\frac{3}{5}+\frac{3}{8}\log_2\frac{3}{5}+\frac{4}{8}\log_2\frac{4}{6}+\frac{4}{8}\log_2\frac{4}{8}+\frac{1}{8}\log_2\frac{1}{5}+\frac{3}{8}\log_2\frac{3}{7}+\frac{0}{8}\log_2\frac{0}{4}]=2.386\ 3,$$

$$HE(D_2\mid B_2)=-[\frac{2}{8}\log_2\frac{2}{6}+\frac{2}{8}\log_2\frac{2}{5}+\frac{2}{8}\log_2\frac{2}{5}+\frac{2}{8}\log_2\frac{2}{6}+\frac{4}{8}\log_2\frac{4}{8}+\frac{4}{8}\log_2\frac{4}{5}+\frac{4}{8}\log_2\frac{4}{7}+\frac{4}{8}\log_2\frac{4}{4}]=2.518\ 1,$$

最后,计算决策划分 U/T_D 关于 B_2 的近似条件熵为：

$$AHE(D\mid B_2)=\log_2(2-\frac{0}{7})\times 2.386\ 3+\log_2(2-\frac{1}{8})\times 2.518\ 1=4.670\ 0$$

基于上述案例机制并经过类似计算,下面直接给出关于属性增链的近似条件熵及其大小关系,即：

$AHE(D\mid B_1)=7.193\ 1>AHE(D\mid B_2)=4.670\ 0>AHE(D\mid B_3)=3.527\ 5>AHE(D\mid B_4)=2.891\ 4>AHE(D\mid C)=1.617\ 6$。

上述关于属性增链的不等式验证了近似条件熵的粒化单调性(定理 2)。所有度量值均隶属理论双界范围 $[0,|U|\log_2|U|]$(推论 1)。

最后根据算法 1 求解该集值决策表的一个属性约简。执行步骤 1,计算决策划分 U/T_D 关于 C 的近似条件熵为 $AHE(D\mid C)=1.617\ 6$。执行步骤 2,设置 $Core_C(D)=\varnothing$,计算 C 中每个属性关于决策划分 U/T_D 的内属性重要度为：

$SIG_{inner}(a_1,C,D)=2.438\ 2-1.617\ 6=0.820\ 6$,

$SIG_{inner}(a_2,C,D)=1.617\ 6-1.617\ 6=0$,

$SIG_{inner}(a_3,C,D)=1.617\ 6-1.617\ 6=0$,

$SIG_{inner}(a_4,C,D)=1.617\ 6-1.617\ 6=0$,

$SIG_{inner}(a_5,C,D)=2.891\ 4-1.617\ 6=1.273\ 8$。

由计算可知,满足条件的属性为 a_1 与 a_5,则 $Core_C(D)=\{a_1,a_5\}$,令 $R=Core_C(D)$。执行步骤 3,计算决策划分 U/T_D 关于 R 的近似条件熵为 $AHE(D\mid R)=3.108\ 6\neq AHE(D\mid C)$。执行步骤 4,$\forall a\in(C-R)$,计算每个属性对于 R 关于决策划分 U/T_D 的外属性重要度为：

$SIG_{outer}(a_2,R,D)=3.108\ 6-2.238\ 6=0.871\ 8$,

$SIG_{outer}(a_3,R,D)=3.108\ 6-2.188\ 1=0.920\ 5$,

$SIG_{outer}(a_4,R,D)=3.108\ 6-1.617\ 6=1.489\ 2$。

选择外属性重要度最大的属性 a_4 并入 R 中,即 R 更新为 $R=\{a_1,a_4,a_5\}$。计算决策划分 U/T_D 关于 R 的近似条件熵为 $AHE(D\mid R)=1.617\ 6=AHE(D\mid C)$。即 R 满足约简第一条,所以执行步骤 5,向前遍历删除 R 中的每个属性 a,有 $AHE(D\mid R-\{a\})\neq AHE(D\mid R)$,进入步骤 6,返回 R,即约简结果为 $R=\{a_1,a_4,a_5\}$。

4　UCI 数据实验

本节实施数据实验来验证近似条件熵及其属性约简,主要是粒化单调性(定理 2)与启发式约简算法(算法 1)。将算法 1 与如下信息角度和代数角度具有代表性的集值决策表启发式约简算法从约简个数与分类精度进行比较,即:基于条件信息熵约简算法[7]、基于信息量约简算法[9]和基于正域约简算法[16]。实验环境为:操作系统 Windows10 64b,Intel(R) Core(TM) i5-8200Y CPU @ 1.61 GHz,内存 4.00 GB,采用 Matlab2018a 进行编程实现。下面从 UCI 数据集中挑选 8 组数据集,见表 2。在实验前均使用文献[16]中的方法,将 UCI 数据集转化为集值决策表。

表 2 数据集

Tab. 2 Data sets

编号	数据集	$\lvert U\rvert$	$\lvert C\rvert$	m
1	Glass	214	9	6
2	Heart	270	13	2
3	Iris	150	4	3
4	Zoo	101	16	7
5	Diabetes	768	8	2
6	Ecoli	336	7	8
7	Vote	435	16	2
8	Vehicle	846	18	4

为表现度量变化,选取自然属性增 $\{a_1\} \subset \{a_1,a_2\} \subset \cdots \subset C$。针对 8 个数据集,分别进行了实验计算,相关度量结果见图 1。在图 1 中, x 轴标识属性个数, y 轴标识近似条件熵度量值。观测可见,近似条件熵值随属性个数的增加而减少,说明其具有属性粒化单调性,与定理 2 一致。基于这些单调曲线,从核出发追求与全部条件属性集近似条件熵相等的属性约简是合理的,则可以得到一个适当长度的约简结果。下面给出 8 个数据集在本文算法下的具体约简结果,见表 3。表 4 则给出 8 个数据集在各约简算法下的约简个数。

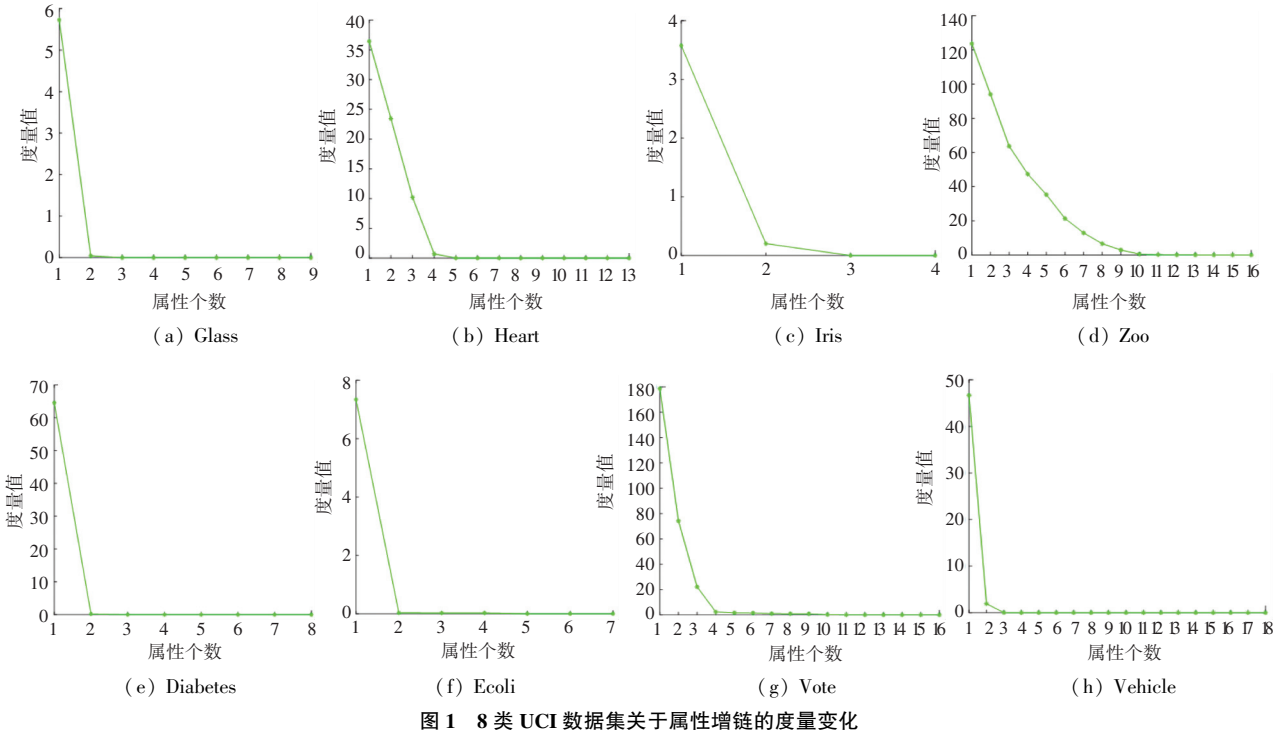

图 1 8 类 UCI 数据集关于属性增链的度量变化

Fig. 1 The measure changes of the 8 types of UCI data sets on the attribute-addition chain

表 3 本文算法的约简结果

Tab. 3 The reduction results of the algorithm in this paper

数据集	约简后的条件属性
Glass	{2,7}
Heart	{1,4,5,8,9}
Iris	{1,3}
Zoo	{4,10,12,13,14}
Diabetes	{1,2,7}
Ecoli	{1,2,5}
Vote	{1,2,3,4,9,11,12,15,16}
Vehicle	{1,4,12}

表 4 不同算法的约简个数

Tab. 4 Number of reductions for different algorithms

数据集	约简后的属性个数			
	文献[7]算法	文献[9]算法	文献[16]算法	本文算法
Glass	3	3	3	2
Heart	5	5	5	5
Iris	3	3	3	2
Zoo	6	5	6	5
Diabetes	3	3	3	3
Ecoli	3	3	3	3
Vote	9	9	9	9
Vehicle	5	3	4	3

由表 4 可知，4 种算法约简后的属性个数都小于原始数据的属性个数，说明对数据集进行优化处理是有必要的。比较 4 种算法的约简个数，本文算法在部分数据集上约简个数更少，例如 Glass、Iris、Zoo 与 Vehicle 数据集；而在其余数据集上，与其它 3 种算法的约简个数相等。说明本文算法的约简结果更优。

最后比较不同算法的分类精度。这里采用 SVM 分类器，对表 4 中 8 个数据集的约简结果进行十折交叉分类训练，得到各算法约简结果的分类精度值，具体见表 5。

表 5　SVM 分类器下不同算法的分类精度比较

Tab. 5　Comparison of classification accuracy of different algorithms under SVM classifier　%

数据集	约简集的分类精度			
	文献[7]算法	文献[9]算法	文献[16]算法	本文算法
Glass	60.7	61.1	60.7	61.6√
Heart	72.2	72.2	72.2	72.2-
Iris	89.3	90.7	89.3	93.3√
Zoo	73.3	78.2√	77.2	78.2√
Diabetes	76.0	76.0	76.0	76.0-
Ecoli	74.7	74.7	74.7	74.7-
Vote	90.3	94.7	94.7	95.4√
Vehicle	65.4	69.7√	68.6	69.7√

在表 5 中，“√”符号标识 4 种算法约简集下分类精度的最大值，“-”符号标识 4 种算法约简集下分类精度持平。观察表 5 可知，本文算法在大部分数据集中，分类精度高于或等于其它 3 种算法，仅在 Heart、Diabetes 与 Ecoli 数据集中与其它 3 种算法持平。这是因为其它 3 种算法都是采用的单一度量方式，存在局限性；而本文算法采用的近似条件熵是由近似精度与条件信息熵信息融合所得，是一种更加全面的度量模型，从而具有更优的分类性能。

5　结束语

本文基于近似精度与条件信息熵，提出近似条件熵，构建基于近似条件熵的属性约简及其启发式约简算法。通过具体实例与数据实验，验证了近似条件熵性质的正确性与约简算法的有效性，实验结果表明，本文算法不仅能够获得更优的约简结果，而且分类精度更高。所得结果深化了信息学习与特征选择，对集值决策表的知识发现与规则推导具有意义。下一步将对集值决策表的规则提取进行研究。此外，降低本文算法的时间复杂度，提高其运行效率，也是接下来的研究重点。

参考文献

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

[2] 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.

[3] 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.

[4] 张宁, 范年柏. 基于邻域近似条件熵的启发式属性约简[J]. 计算机应用研究, 2018, 35(5): 1395-1398.

[5] 李晨光, 乔帅, 杨晓杰, 等. 基于 KNE-BPNN 的电务设备故障预测[J]. 计算机应用研究, 2019, 36(9): 2712-2717.

[6] 牟恩, 张贤勇, 姚岳松, 等. 邻域近似条件熵的特定类属性约简及启发算法[J]. 计算机工程与应用, 2020, 56(24): 175-180.

[7] 苏莉. 基于条件信息熵的多值信息系统的属性约简[J]. 电脑学习, 2010(6): 106-107.

[8] DAI Jianhua, TIAN Haowei. Entropy measures and granularity measures for set-valued information systems[J]. Information Sciences, 2013, 240: 72-82.

[9] 马建敏, 张文修. 基于信息量的集值信息系统的属性约简[J]. 模糊系统与数学, 2013, 27(2): 177-182.

[10] 庄颖, 刘文奇, 范敏, 等. 集值信息系统上的多粒度优势关系与信息融合[J]. 模式识别与人工智能, 2015, 28(8): 741-749.

[11] LIU Yi, ZHONG Chunzhen. Attribute reduction of set-valued decision information system based on dominance relation[J]. Journal of Interdisciplinary Mathematics, 2016, 19(3): 469-479.

[12] WEI Wei, CUI Junbiao, LIANG Jiye, et al. Fuzzy rough approximations for set-valued data[J]. Information Sciences, 2016, 360(C): 181-201.

[13] QIAN Wenbin, SHU Wenhao, ZHANG Changsheng. Feature selection from the perspective of knowledge granulation in dynamic set-valued information system[J]. Journal of Information Science and Engineering, 2016, 32(3): 783-798.

[14] HUANG Yanyong, LI Tianrui, LUO Chuan, et al. Dynamic variable precision rough set approach for probabilistic set-valued information systems[J]. Knowledge-Based Systems, 2017, 122(5): 131-147.

[15] 王映龙, 华佳佳, 钱文彬, 等. 集值决策信息系统的增量式属性约简算法[J]. 小型微型计算机系统, 2018, 39(6): 1239-1244.

[16] 陈曼如, 张楠, 童向荣, 等. 集值信息系统的快速正域约简[J]. 智能系统学报, 2019, 14(3): 471-478.

[17] SINGH S, SHREEVASTAVA S, SOM T, et al. A fuzzy similarity-based rough set approach for attribute selection in set-valued information systems[J]. Soft Computing, 2020, 24(6): 4675-4691.

[18] 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.

[19] 唐鹏飞. 区间集决策表不确定性度量的修正 δ-区间决策条件熵方法[J]. 内江师范学院学报, 2021, 36(6): 34-39.