山西大学学报(自然科学版)46(1):101－107,2023
Journal of Shanxi University(Nat.Sci.Ed.)

DOI:10.13451/j.sxu.ns.2022107

区间值序决策表的条件熵属性约简

张晓燕*,匡洪毅
(西南大学 人工智能学院,重庆 400715)

摘　要:由于数据自身的不确定性和观测条件有限,现实问题中许多数据以区间值形式呈现。其中,优势关系下的区间值信息表研究对于多属性决策问题有重要意义。目前针对该系统的属性约简方法主要是辨识矩阵法或基于互信息的增量式约简,但前者计算效率较低,而后者没有利用到决策信息。文章探讨了条件熵作为不确定性度量在该系统下的性质,通过比较不同属性缺失时信息系统的条件熵变化,引入了属性重要度概念,基于此提出启发式属性约简算法。最后,通过对比实验验证了本算法具有低冗余的特点,在约简率上比基于粗糙熵和正域不变等序信息系统的启发式约简。

关键词:区间值序决策表;条件熵;属性约简;属性重要度
中图分类号:TP18　　**文献标志码**:A　　**文章编号**:0253-2395(2023)01-0101-07

Attribute Reduction Based on Conditional Entropy in Interval Valued Ordered Decision Table

ZHANG Xiaoyan*, KUANG Hongyi
(*College of Artificial Intelligence, Southwest University, Chongqing* 400715*, China*)

Abstract: Because of the uncertainty of the data and the conditions of limited observation, many practical problems are presented as interval valued data. Among them, the study of interval-valued information system based on the dominance relationship has important significance for multi-attribute decision-making. At present, the main attribute reduction methods for this system are discernibility matrix method or incremental reduction based on mutual information, but the former is computationally inefficient, while the latter ignores decision information. This paper discusses the properties of conditional entropy as a measure of uncertainty in this system. By comparing the changes of conditional entropy of information system when different attributes are missing, the concept of attribute importance is introduced, and a heuristic attribute reduction algorithm is proposed based on this. Finally, the comparative experiment verifies that the algorithm has the characteristics of low redundancy, and the reduction rate is 14%-25% higher than that of heuristic reduction based on rough entropy and positive region-based in ordered information system.

Key words: interval-value ordered decision table; conditional entropy; attribute reduction; attribute importance

0　引言

粗糙集理论[1]最初由波兰数学家Pawlak提出,是一种处理不精确、不一致与不完全数据的数学工具。近年来,随着知识发现、机器学习、数据挖掘、决策分析等领域的兴起,粗糙集理论的研究与应用也逐渐得到更多学者的关注[2]。

收稿日期:2022-08-15;**接受日期**:2022-10-25
基金项目:国家自然科学基金(61976245)
* **通信作者**:张晓燕(1979-),女,山西怀仁人,博士,副教授,主要研究方向为人工智能的数学基础、形式概念分析、粒计算、知识发现、不确定性处理等。E-mail:zxy19790915@163.com
引文格式:张晓燕,匡洪毅.区间值序决策表的条件熵属性约简[J].山西大学学报(自然科学版),2023,46(1):101-107.
DOI:10.13451/j.sxu.ns.2022107

现实问题中,由于环境的复杂性(噪音、信息缺失、决策者具有偏好特性等),许多问题的描述不是一个精确的值,而很可能是一个模糊的数值区间。同时,数据的属性值之间往往存在大小、强弱等优势关系,因此优势关系下的粗糙集模型相对于经典模型具有更广阔的应用空间。例如,多属性决策问题中需要对样本的属性值进行排序从而筛选出最优方案。因此,对于连续且具有偏序关系的区间值数据的研究具有重要价值。

知识约简是粗糙集理论的核心问题之一。所谓知识约简,就是在保持信息系统的分类能力不变的前提下,删除其中的冗余属性,从而快速地从庞大、繁杂的数据中提取有效信息。现有的属性约简方法可分为以下几类:基于差别矩阵的属性约简算法[3]、基于正区域的属性约简算法[4-5]以及基于信息论的约简算法[6]。

针对区间值序信息系统的约简方法研究中,Qian等[7]最早提出一种通过比较区间的上下边界给对象排序的方法,并给出了基于优势粗糙集(DRSA)的属性约简方法。Shi等[8]对区间值进行模糊化,并求其分布约简判定条件和可辨识矩阵。Yang等[9]为减少差别矩阵的冗余元素,提出差别信息树并提高了计算效率。Zhang等[10]基于α-容差关系,保证系统的上下近似在特定意义下不变,提出了α-上、下近似约简。上述方法均通过构建差别矩阵进行约简,其时空复杂度较高。而启发式属性约简算法是一种更高效的约简途径,即通过构建评估属性重要程度的指标并结合贪心算法得到约简。其中,利用信息论的观点评估属性重要度是一种不错的方法。Dai等[11]提出了基于α-弱相似的不完备区间值信息系统的不确定性度量方法。Feng等[12]基于覆盖诱导的划分,构造了区间值信息系统的信息熵和补熵度量方式。Yan等[13]从互信息的角度讨论了特征重要性,并以此构建增量式约简算法。

本文在上述研究的基础上,针对优势关系下的区间值决策表作进一步研究。文献[13]可作为一种无监督的特征选择方法,但当数据包含决策信息时,将其考虑在内能更好地提取到有效信息。因此,本文讨论了条件熵作为不确定性测度的可行性,并基于此提出了启发式属性约简算法。最后,通过对比实验验证了本算法在约简率上的优越性。

1 预备知识

1.1 粗糙集理论与序信息表

信息系统又称信息表或知识表示系统,主要表现形式是一张反映对象与属性之间关系的数据表。一张信息表可形象地表示为$I=(U,AT,V,f)$,其中,U表示对象集论域,AT是属性集合,V是属性值域,$f:U\times AT\to V$是信息函数,表示论域到属性集的映射。Pawlak粗糙集在近似空间下为信息表赋予了新意义,它使得每个属性集确定一个二元不可区分关系(等价关系)。但在实际问题中,并非所有信息系统都是可精确划分的,有不少信息系统是基于偏序关系,也称优势关系。针对这种情况,序信息系统应运而生,下面给出相关定义。

在一个信息系统中,如果在某属性值域上存在优势关系,则称这个属性为一个准则。当所有的属性都为准则时,该信息系统称为序信息系统[14],形象地表示为$I^{\geqslant}=(U,AT,V,f)$。对$\forall x_i,x_j\in U$,可用"$x_i\geqslant_a x_j$"表示x_i在准则a下优于x_j。若在属性集$A\subseteq AT$下比较两对象,则"$x_i\geqslant_A x_j$"表示x_i相对于A中的所有属性都优于或者等于x_j。

设$I^{\geqslant}=(U,AT,V,f)$为一序信息系统,类似于粗糙集中的等价关系,并基于上述描述,可定义序信息系统中的优势关系$R_A^{\geqslant}$为:

$$R_{AT}^{\geqslant}=\{(x,y)\in U\times U;f_a(x)\geqslant f_a(y),\forall a\in AT\}。$$

由条件属性上的优势关系所诱导的优势类$[x_i]_A^{\geqslant}$及覆盖$U/R_A^{\geqslant}$可分别表示为:

$$[x_i]_A^{\geqslant}=\{x_j\in U\mid(x_i,x_j)\in R_A^{\geqslant}\}\Leftrightarrow$$
$$\{x_j\in U\mid(\forall a\in A)[f_a(x_i)\leqslant f_a(x_j)]\}。$$

由条件属性集AT和决策属性集D所诱导的覆盖簇分别为

$$U/R_{AT}^{\geqslant}=\{[x]_A^{\geqslant}\mid x\in U\},$$
$$U/R_D^{\geqslant}=\{D_1,D_2,\cdots,D_s\}。$$

1.2 区间值序决策表

在现实生活中,并非所有问题的属性都可

以被刻画为一个确切的值,而很可能是一个不确定的数值区间或范围。对拥有这种属性的信息表定义如下[15]。

设$I=(U,AT\cup D,F,G)$为一个决策信息表,若它满足对$\forall x\in U,\forall a\in AT,\forall f\in F$,则有:

$$f(x_i,a)=\left[a^L(x_i),a^U(x_i)\right],$$

称其为区间值决策表,其中$f(x_i,a)$表示x_i在属性a下的属性值区间,$a^L(x_i)$和$a^U(x_i)$分别为区间的左右端点,且$a^L(x_i)\leqslant a^U(x_i)$。

对具有优势关系的区间值信息表可以表示为$I=(U,AT\cup D,F,G)$,对准则$a\in AT$所诱导的优势关系表示为

$$f(x_i,a)\leqslant f(x_j,a)\Leftrightarrow$$
$$\left[a^L(x_i)\leqslant a^L(x_j)\wedge a^R(x_i)\leqslant a^R(x_j)\right]。$$

若对$\forall a\in AT$都满足上述优势关系,则称该区间值信息表为区间值序决策表,其中,由优势关系集合$A\in AT$所诱导的优势类和覆盖为

$$[x_i]_A^{\geqslant}=R_A^{\geqslant}(x_i)=$$
$$\left\{x_j\in U\,\middle|\,(\forall a\in A)\left[a^L(x_i)\leqslant a^L(x_j)\wedge a^R(x_i)\leqslant a^R(x_j)\right]\right\}$$

为了更直观方便地理解,下面给出区间值序决策表的完整数学定义。区间值序决策表可以用一个五元组表示为$I^{\geqslant}=(U,AT\cup D,F,G)$,其中$U=\{u_1,u_2,\cdots,u_n\}$,表示有限对象集论域;$AT=\{a_1,a_2,\cdots,a_T\}$,表示有限条件属性集;$D=\{d_1,d_2,\cdots,d_s\}$,表示有限决策属性集合;$F=\left\{f:U\to V_a,a\in AT\right\}$,表示论域与条件属性集的关系集合;$G=\left\{g:U\to V_d,d\in D\right\}$,表示论域与决策属性集的关系集合。

设$I=(U,AT\cup D,F,G)$为区间值序决策表,易证其上的优势关系$R_A^{\geqslant}$有如下性质:

(1)自反性、传递性,但不一定具有对称性;

(2)若$P\subseteq Q\subseteq AT$,则:$R_Q^{\geqslant}\subseteq R_P^{\geqslant}$;

(3)若$P\subseteq Q\subseteq AT$,则:$[u_i]_Q^{\geqslant}\subseteq[u_i]_P^{\geqslant}$。

定义1 类似于粗糙集中等价关系的上下近似定义,在区间值序决策表中,对优势关系也可定义一对上下近似算子。

$$\underline{R_B^{\geqslant}}(X)=\left\{u_i\in X\,|\,[u_i]_B^{\geqslant}\subseteq X\right\},$$
$$\overline{R_B^{\geqslant}}(X)=\left\{u_i\in X\,|\,[u_i]_B^{\geqslant}\cap X\neq\Phi\right\}。$$

若$R_{AT}^{\geqslant}\subseteq R_D^{\geqslant}$,则称该系统为协调的一致的,否则为不协调的或不一致的。本文就不协调的信息系统进行研究。

2 知识的粗糙条件熵与属性重要度及属性约简算法

2.1 区间值序决策表中知识的粗糙条件熵

属性对系统的协调性贡献是评价属性重要性的关键指标,即去掉某一属性,若系统不协调程度(也称粗糙度)增加越多,则该属性越重要。而条件熵就可作为该指标,其主要思想是:在已知某变量的情况下,评估另一变量的不确定性。在区间值序决策表中,我们引入该概念来评价此类信息系统的粗糙程度。

定义2 设$I^{\geqslant}=(U,AT\cup D,F,G)$为一区间值序决策表,在属性集$A\subseteq AT$下,粗糙条件熵定义为

$$H(D|A)=-\sum_{i=1}^{|U|}\sum_{j=1}^{m}\frac{\left|R_A^{\geqslant}(u_i)\cap D_j\right|}{|U|}\log\frac{\left|R_A^{\geqslant}(u_i)\cap D_j\right|}{R_A^{\geqslant}(u_i)}。$$

定理1**(最大值)** 设$I^{\geqslant}=(U,AT\cup D,F,G)$为一区间值序决策表,且$P\subseteq AT$,$I^{\geqslant}$的粗糙条件熵最大值为$U\log|U|$,当且仅当$\forall u_i\in U,R_P^{\geqslant}(u_i)=U$且$\forall D_i\in U/D,\left|D_j\right|=1$,此时优势关系退化等价关系。

定理2**(最小值)** 设$I^{\geqslant}=(U,AT\cup D,F,G)$为一区间值序决策表,且$P\subseteq AT$,$D$相对于$P$的粗糙条件熵最小值为0,当且仅当对$\forall d_i,d_j\in D$,有$d_i=d_j$。

定理3**(单调性)** 设$I^{\geqslant}=(U,AT\cup D,F,G)$为一区间值序决策表,且$P\subseteq AT,Q\subseteq AT$若$P\subseteq Q$,则$H(D|Q)\leqslant H(D|P)$。

证明 当$P\subseteq Q$时,对$u_i\in U$,有$R_Q^{\geqslant}(u_i)\subseteq R_P^{\geqslant}(u_i)$,则$\left|R_Q^{\geqslant}(u_i)\right|\subseteq\left|R_P^{\geqslant}(u_i)\right|$。首先定义如下概念

$$R_{Q_{i,x}}^{\geqslant}(u_i)=R_Q^{\geqslant}(u_i)\cap D_j,$$
$$R_{Q_{i,y}}^{\geqslant}(u_i)=R_Q^{\geqslant}(u_i)\cap(U-D_j)。$$

若用$x_Q^{i,j}$替代$\left|R_{Q_{i,x}}^{\geqslant}(u_i)\right|$,$y_Q^{i,j}$替代$\left|R_{Q_{i,y}}^{\geqslant}(u_i)\right|$,则:$R_{Q_j}^{\geqslant}(u_i)=x_Q^{i,j}+y_Q^{i,j}$。其中,$R_{Q_j}^{\geqslant}(u_i)$表示$R_Q^{\geqslant}(u_i)$与决策类$D_j$的交集,同理对属性集$P$也有上述定义。显然地有,$x_P^{i,j}\leqslant x_Q^{i,j}$,$y_P^{i,j}\leqslant y_Q^{i,j}$,继而对条件熵公式可做如下变换:

$$H(D|A)=$$

$$-\sum_{i=1}^{|U|}\sum_{j=1}^{m}\frac{\left|R_A^{\geqslant}(u_i)\cap D_j\right|}{|U|}\log\frac{\left|R_A^{\geqslant}(u_i)\cap D_j\right|}{\left|R_A^{\geqslant}(u_i)\right|}=-\sum_{i=1}^{|U|}\sum_{j=1}^{m}\frac{x_A^{i,j}}{|U|}\log\frac{x_A^{i,j}}{x_A^{i,j}+y_A^{i,j}}。$$

对求和的每一项可看作一个二元函数

$$f(x,y)=-x\log(x/(x+y))(x>0,y\geqslant 0),$$

分别对x和y求偏导都可得到单调递增函数，因此

$$f(x_P^{i,j},y_P^{i,j})\leqslant f(x_P^{i,j}+y_Q^{i,j})\leqslant f(x_Q^{i,j}+y_Q^{i,j}),$$

即：$H(D|Q)\leqslant H(D|P)$。该定理说明随着分辨能力的增强，粗糙条件熵单调递减，即决策属性相对于条件属性的不确定性增加、系统更加粗糙。

2.2 区间值序决策表中属性的重要度

本节给出基于粗糙条件熵的属性重要度判定方法，作为该系统下启发式属性约简算法的基础。

定义3 设$I^{\geqslant}=(U,AT\cup D,F,G)$为区间值序决策表，且$A\subseteq AT$，定义属性$a$在$A$中的绝对重要度为

$$DS(a,A)=H(D|[A\backslash a])-H(D|A)。$$

记属性集A的核为$Core(A)$，表示刻画决策属性所必要的属性。

$$Core(A)=\{a\in A|DS(a,A)>0\}。$$

定理4 结合上述定义及定理1和定理2易得

(1) $0\leqslant DS(a,A)\leqslant|U|\log|U|$；

(2)当$A=\{a\}$时，用$DS(a)$替代$DS(a,A)$，$DS(a)=|U|\log|U|-H(D|\{a\})$。

定义4 设$I^{\geqslant}=(U,AT\cup D,F,G)$为区间值序决策表，$C\subseteq A$，对$\forall a\in A\backslash C$，定义$a$相对于$C$的相对重要度为

$$DR(a,C)=H(D|C)-H(D|[C\cup a])。$$

由上述定义可知，当$DR(a,C)$值越大时，属性a相对于C的重要度就越高。因此在属性约简过程中，每一步都先将满足

$$H(D|C\cup\{a\})=\min_{a'\in AT-C}\left\{H(D|C\cup\{a'\})\right\}$$

的属性纳入核中，得到次最优或最优约简。

定义5 设$I^{\geqslant}=(U,AT\cup D,F,G)$为区间值序决策表，$B\subseteq AT$，若满足$H(D|B)=H(D|AT)$，且对$\forall a\in B$，有$DR(a,B)>0$，则$B$为$AT$的一个约简。由定义3可求得属性集$AT$的核，由于核唯一并且为任何约简的子集，因此核可以作为最小约简的起点。由定义4中的属性重要度，基于当下未选的属性集合，逐步将最重要的属性添加到核中，直至其粗糙条件熵等于$H(D|AT)$。即在$Core(AT)$的基础上通过增加属性构成的最小约简。

2.3 基于条件熵的区间值序决策表属性约简算法

首先，对给定的一个区间值序信息系统，根据定义3计算属性集的核。其次，将属性集合分为两部分，一部分为已选择属性，即约简属性集，另一部分为未选择属性，以核作为约简的起点。接着，根据定义4中的属性重要度，对未选择的属性集排序，选择对于约简集最重要的属性并添加到约简集中。然后，更新上述两部分集合并重复上一步操作。最后，当约简集的条件熵等于原信息表的条件熵时完成约简。算法的具体实现见表1。

表1 基于条件熵的区间值序决策表属性约简算法

Table 1 Algorithm of attribute reduction based on conditional entropy in interval valued ordered decision table

算法1 区间值序决策表下基于条件熵的启发式属性约简
输入：区间值序决策$I^{\geqslant}=(U,AT\cup D,F,G)$。
输出：该区间值序决策表的约简。
1: $Core(AT)\leftarrow\{a\in A|DS(a,A)>0\}$
2: $Red(AT)\leftarrow Core(AT)$
3: *while* $H(D|Red(AT))\neq H(D|AT)$ *do*
4: $S\leftarrow AT-Red(AT)$
5: $\max DR=0$
6: *for each* $a\in S$:
7: *if* $DR(a,Red(AT))>\max DR$ *then*
8: $\max DR\leftarrow DR(a,Red(AT))$
9: $\max\_a\leftarrow a$
10: *end*
11: *end*
12: $Red(AT)\leftarrow Red(AT)\cup\{\max\_a\}$
13: *end*
14: *return Red(A)*

注：表中$Core(*)$表示核属性，$Red(*)$表示当前约简，$H(*)$表示条件熵。

下面给出算法1时间复杂度分析。首先令$|U|=n$，$|AT|=r$，$|D|=s$，$|Core(AT)|=r_1$，$|Red(AT)|=r_2$。为方便计算，对$\forall A\subseteq AT$，先讨论其条件熵计算的时间复杂度，记$O(\beta)=$

表2 算法1的时间复杂度分析

Table 2 Analysis of the time complexity of Algorithm 1

步骤1	$O(r\times O(\beta))$
步骤6—11	$O(r\times O(\beta))$
步骤3—13	$O(r\times(r_2-r_1)\times O(\beta))$
总计	$O(r\times O(\beta)+r\times(r_2-r_1)\times O(\beta))$

$O(H(D|A))$。对单个属性下,优势类的计算可采用0—1矩阵,则属性集的优势类矩阵为各属性下的优势矩阵求交集,其计算复杂度为$O(n^2\times r)$。进而结合定义2,易得$O(\beta)=O(n^2\times r+n\times s)$。于是可得算法1的详细时间复杂度,如表2所示。

3 案例分析

下面给出一个具体案例来说明本文给出方法的具体操作步骤。设$I^{\geqslant}=(U,AT\cup D,F,G)$为一个区间值序决策表,论域代表6个投资对象:$U=\{u_1,u_2,u_3,u_4,u_5,u_6\}$,属性集代表3种不同的风险:$AT=\{a_1,a_2,a_3\}$,属性值代表风险范围,决策属性$d=\{1,2,3\}$代表风险等级。统计数据如表3所示。

表3 风险投资案例的区间值序决策表

Table 3 Interval-valued ordered decision table for a venture capital case

U	a_1	a_2	a_3	d
u_1	[0.1, 0.3]	[0.2, 0.3]	[0.1, 0.4]	3
u_2	[0.3, 0.5]	[0.2, 0.6]	[0.2, 0.8]	2
u_3	[0.1, 0.5]	[0.1, 0.4]	[0.2, 0.7]	1
u_4	[0.2, 0.7]	[0.1, 0.5]	[0.3, 0.7]	2
u_5	[0.3, 0.6]	[0.3, 0.7]	[0.2, 0.9]	3
u_6	[0.3, 0.9]	[0.2, 0.7]	[0.3, 0.8]	1

注:例如$[u_1,a_1]$所对应的[0.1, 0.3]表示投资对象u_1在a_1方面的风险值所在范围。

对于表3,根据决策属性d的不同,可将论域划分为三类:$D_1=\{x_1,x_5\}$,$D_2=\{x_2,x_4\}$,$D_3=\{x_3,x_6\}$。

原表下各对象的优势类如下:

根据属性约简算法,对表3进行属性约简。首先求条件属性集的核$Core(AT)$。由定义2

$$[x_1]_{AT}^{\geqslant}=\{x_1,x_2,x_5,x_6\},$$
$$[x_2]_{AT}^{\geqslant}=\{x_2,x_5,x_6\},$$
$$[x_3]_{AT}^{\geqslant}=\{x_2,x_3,x_4,x_5,x_6\},$$
$$[x_4]_{AT}^{\geqslant}=\{x_4,x_6\},$$
$$[x_5]_{AT}^{\geqslant}=\{x_5\},[x_6]_{AT}^{\geqslant}=\{x_6\},$$

可算得原信息表的条件熵$H(D|AT)\approx 3.3941$。

若去掉属性a_1,容易验证在$C_1=\{a_2,a_3\}$下,对$\forall x\in U$有$[x]_{AT}^{\geqslant}=[x]_{C_1}^{\geqslant}$。由条件熵公式可知,条件熵不变,即$DS(a_1,AT)=0$。

若去掉属性a_2,在$C_2=\{a_1,a_3\}$下,论域的优势类为:

$$[x_1]_{C_2}^{\geqslant}=\{x_1,x_2,x_3,x_4,x_5,x_6\},$$
$$[x_2]_{C_2}^{\geqslant}=\{x_2,x_5,x_6\},$$
$$[x_3]_{C_2}^{\geqslant}=\{x_2,x_3,x_4,x_5,x_6\},$$
$$[x_4]_{C_2}^{\geqslant}=\{x_4,x_6\},$$
$$[x_5]_{C_2}^{\geqslant}=\{x_5\},[x_6]_{C_2}^{\geqslant}=\{x_6\}。$$

$H(D|AT/a_2)\approx 3.9791>H(D|AT)$。因此满足$DS(a_2,AT)>0$,将$a_2$纳入核中。同理,若去掉$a_3$,$C_3=\{a_1,a_2\}$的情况如下:

$$[x_1]_{C_3}^{\geqslant}=\{x_1,x_2,x_5,x_6\},$$
$$[x_2]_{C_3}^{\geqslant}=\{x_2,x_5,x_6\},$$
$$[x_3]_{C3}^{\geqslant}=\{x_2,x_3,x_4,x_5,x_6\},$$
$$[x_4]_{C_3}^{\geqslant}=\{x_4,x_6\},$$
$$[x_5]_{C_3}^{\geqslant}=\{x_5\},[x_6]_{C_3}^{\geqslant}=\{x_6\}。$$

$H(D|AT/a_3)\approx 3.3941>H(D|AT)$,所以$a_3$也不是核属性,则$Core(AT)=\{a_2\}$。而$H(D|\{a_1,a_2\})=H(D|\{a_2,a_3\})=H(D|AT)$已被验证,因此$\{a_1,a_2\}$和$\{a_2,a_3\}$都是原属性集的约简。

4 实验分析

本节将通过实验对2节中提出的算法性能进行验证,并与另外两个算法分析比较。本文从UCI数据集网站 http://archive.ics.uci.edu/ml/datasets.html 下载了6个数据集,数据的具体信息如表4所示。整个实验在一台私人电脑上实现,其具体配置如表5所示。

表4 实验所用数据集信息

Table 4 Dataset information used in the experiment

名称	样本数	特征数	类别数	总元素个数
Glass	214	9	7	2140
Autism	704	20	2	14 784
Wdbc	569	30	2	18 476
German	1000	24	2	25 000
Wine	4898	11	6	58 776
Segmentation	6435	36	7	238 095

表5 实验运行环境信息

Table 5 Information of operating environment

名称	型号	参数
CPU	AMD Ryzen 7 4800H	2.90 GHz
软件	Python	3.9
系统	Windows 10	64 bit
内存	SAMSUNG DDR4	16 GB; 2666 MHz
硬盘	Intel SSDPEKNW	512 GB

首先,将数据集转化为区间值形式,具体做法[16]为:假设实值数据集$(U, C\cup\{d\}, V, f)$,$a\in C$,则

$$a_i^- = a_i - 2\sigma_d, a_i^- = a_i + 2\sigma_d,$$

其中,对样本x_i来说,σ_d为与x_i同标签的样本在属性a下特征值的标准差。

然后,将算法1和对比算法分别对6个区间值序信息系统进行属性约简。对比算法如下:(1)文献[15]中针对序信息系统提出了一种基于粗糙熵(类似于信息熵)的启发式属性约简方法,我们将其引入到区间值序信息系统中;(2)文献[16]针对区间值信息系统提出了一种基于正域不变的启发式属性约简方法,同样也可拓展到本系统。上述方法的约简思路与本文相似,但衡量属性重要程度的标准不同。

最后,从属性约简率和约简后信息系统的分类能力两方面进行对比。

表6 3种算法下数据集属性的约简率

Table 6 The reduction rate of dataset attributes under three algorithms

数据集	约简前特征数	条件熵		粗糙熵		正域	
		N	R	N	R	N	R
Glass	9	5	44%	6	33%	6	33%
Autism	20	2	90%	18	10%	8	60%
Wdbc	30	4	86%	4	86%	5	83%
German	24	3	88%	10	58%	3	88%
Wine	11	6	46%	9	18%	9	18%
Sege	36	5	86%	6	83%	10	72%

注:表中N、R分别表示约简后的特征数和约简率。

表7 3种算法约简后数据集的KNN分类效果

Table 7 KNN classification effect of three algorithms

数据集精度	条件熵	粗糙熵	正域
Glass	0.640 8	0.509 2	0.677 8
Autism	0.811 4	0.942 8	0.988 5
Wdbc	0.929 5	0.809 4	0.950 6
German	0.784 9	0.724 0	0.735 0
Wine	0.556 1	0.540 8	0.563 2
Seg	0.930 9	0.921 4	0.659 5

关于分类器的选择,目前可用的区间值数据分类器很少,采用文献[17]中扩展的K近邻(KNN)分类器来比较分类效果。对于每种算法,随机抽取10%的样本,代入不同K值($1\leqslant K\leqslant\sqrt{N}$,$N$为样本数)来选取该算法适合的$K$值,采用十折交叉验证做分类预测。属性约简率如表6所示,分类效果结果如表7所示。

从表6可以得出,条件熵的属性约简率均高于另外两种算法。相较于粗糙熵和正域的约简算法,本文算法的约简性能更稳定、无低约简率出现,使得所有数据集均得到有效约简,平均约简率达73.3%,高于另外两种算法14%~25%。

从表7的约简后分类效果来看,条件熵算法的分类效果大多优于粗糙熵算法并与保持正域算法持平,且总体分类效果较好。综合来看,相较于另外两种算法,条件熵算法能保持高约简率的同时分类效果较好。因此,本文提出的算法在区间值序信息系统下是可行的且具有应用价值。

5 结论

本文在区间值信息系统下进行研究,基于信息论的中条件熵概念提出了一种搜索式属性约简算法。条件熵可以充分利用数据的标签信息,以此评价数据中支持正确分类的信息量,并筛选掉对分类不重要的属性。基于条件熵,定义了属性重要度并提出启发式约简算法,以确保我们逐步选取的属性都是当前最优解,从而构成最佳约简。之后,分析了算法的时间复杂度。最后,通过实验分析并与一些该信息系统下的约简算法做对比,验证了该算法的有效性。在未来的工作中,将针对区间值序信息系统下不同的偏序定义进行研究,并将该算法推广到其中。

参考文献:

[1] PAWLAK Z, GRZYMALA-BUSSE J, SLOWINSKI R, *et al*. Rough Sets[J]. *Commun ACM*, 1995, **38**(11): 88−95. DOI: 10.1145/219717.219791.

[2] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, **32**(7): 1229−1246. DOI: 10.3724/SP.J.1016.2009.01229.
WANG G Y, YAO Y Y, YU H. A Survey on Rough Set Theory and Applications[J]. *Chin J Comput*, 2009, **32**(7): 1229−1246. DOI: 10.3724/SP.J.1016.2009.01229.

[3] JELONEK J, KRAWIEC K, SLOWIŃSKI R. Rough Set Reduction of Attributes and Their Domains for Neural Networks[J]. *Comput Intell*, 1995, **11**(2): 339−347. DOI: 10.1111/j.1467-8640.1995.tb00036.x.

[4] 刘少辉, 盛秋戬, 吴斌, 等. Rough集高效算法的研究[J]. 计算机学报, 2003, **26**(5): 524−529. DOI: 10.3321/j.issn: 0254-4164.2003.05.002.
LIU S H, SHENG Q J, WU B, *et al*. Research on Efficient Algorithms for Rough Set Methods[J]. *Chin J Comput*, 2003, **26**(5): 524−529. DOI: 10.3321/j. issn: 0254-4164. 2003.05.002.

[5] GUAN J W, BELL D A. Rough Computational Methods for Information Systems[J]. *Artif Intell*, 1998, **105**(1/2): 77−103. DOI: 10.1016/S0004-3702(98)00090-3.

[6] 刘振华, 刘三阳, 王珏. 基于信息量的一种属性约简算法[J]. 西安电子科技大学学报, 2003, **30**(6): 835−838. DOI: 10.3969/j.issn.1001-2400.2003.06.028.
LIU Z H, LIU S Y, WANG J. An Attribute Reduction Algorithm Based on the Information Quantity[J]. *J Xidian Univ*, 2003, **30**(6): 835−838. DOI: 10.3969/j. issn. 1001-2400.2003.06.028.

[7] QIAN Y H, LIANG J Y, DANG C Y. Interval Ordered Information Systems[J]. *Comput Math Appl*, 2008, **56**(8): 1994−2009. DOI: 10.1016/j.camwa.2008.04.021.

[8] 史德容, 徐伟华. 区间值模糊决策序信息系统的分布约简[J]. 计算机科学与探索, 2017, **11**(4): 652−658. DOI: 10.3778/j.issn.1673-9418.1602002.
SHI D R, XU W H. Distribution Reduction in Interval-valued Fuzzy Decision Ordered Information Systems[J]. *J Front Comput Sci Technol*, 2017, **11**(4): 652−658. DOI: 10.3778/j.issn.1673-9418.1602002.

[9] 杨蕾, 张晓燕, 徐伟华. 区间值序信息系统中差别信息树的属性约简[J]. 计算机科学与探索, 2019, **13**(6): 1062−1069 DOI: 10.3778/j.issn.1673-9418.1805037.
YANG L, ZHANG X Y, XU W H. Attribute Reduction of Discernibility Information Tree in Interval-valued Ordered Information System[J]. *J Front Comput Sci Technol*, 2019, **13**(6): 1062−1069. DOI: 10.3778/j.issn.1673-9418. 1805037.

[10] 张晓雨, 李同军. 不协调区间值决策系统中的α-上,下近似约简[J]. 山东大学学报(理学版), 2022, **57**(5): 20−27. DOI: 10.6040/j.issn.1671-9352.7.2021.216.
ZHANG X Y, LI T J. α-lower and Upper Approximation Reductions in Inconsistent Interval-vlued Decision Systems [J]. *J Shandong Univ Nat Sci*, 2022, **57**(5): 20−27. DOI: 10.6040/j.issn.1671-9352.7.2021.216.

[11] DAI J H, WEI B J, ZHANG X H. *et al*.. Uncertainty Measurement for Incomplete Interval-valued Information Systems Based on α-weak Similarity[J]. *Knowl Based Syst*, 2017, **136**: 159−171. DOI: 10.1016/j.knosys. 2017. 09.009.

[12] 冯琴荣, 温玮华. 区间值信息系统的熵度量[J]. 电子科技大学学报, 2021, **50**(1): 101−105. DOI: 10.12178/1001-0548.2019243.
FENG Q R, WEN W H. Entropy Measurement for Interval-Valued Information Systems[J]. *J Univ Electron Sci Technol China*, 2021, **50**(1): 101−105. DOI: 10.12178/1001-0548.2019243.

[13] 闫岳君, 代建华. 区间序信息系统的无监督特征选择[J]. 模式识别与人工智能, 2017, **30**(10): 928−936. DOI: 10.16451/j.cnki.issn1003-6059.201710007.
YAN Y J, DAI J H. Unsupervised Feature Selection for Interval Ordered Information Systems[J]. *Pattern Recognit Artif Intell*, 2017, **30**(10): 928−936. DOI: 10.16451/j.cnki.issn1003-6059.201710007.

[14] 徐伟华. 序信息系统与粗糙集介绍及研究综述[J]. 琼州学院学报, 2014, **21**(5): 12−16. DOI: 10.13307/j.issn. 1008-6722.2014.05.03.
XU W H. Review of Ordered Information Systems and Rough Set Theory[J]. *J Qiongzhou Univ*, 2014, **21**(5): 12−16. DOI: 10.13307/j.issn.1008-6722.2014.05.03.

[15] 徐伟华, 张晓燕, 钟坚敏, 等. 序信息系统中属性约简的启发式算法[J]. 计算机工程, 2010, **36**(17): 69−71. DOI: 10.3969/j.issn.1000-3428.2010.17.024.
XU W H, ZHANG X Y, ZHONG J M, *et al*. Heuristic Algorithm for Attributes Reduction in Ordered Information Systems[J]. *Comput Eng*, 2010, **36**(17): 69−71. DOI: 10.3969/j.issn.1000-3428.2010.17.024.

[16] 陈华峰, 龙建武, 瞿先平. 区间值决策信息系统中基于正域的属性约简[J]. 重庆理工大学学报(自然科学), 2019, **33**(11): 130−136. DOI: 10.3969/j. issn. 1674-8425(z). 2019.11.019.
CHEN H F, LONG J W, QU X P. A Positive Region-based Attribute Reduction Approach in Interval-valued Decision Information System[J]. *J Chongqing Univ Technol Nat Sci*, 2019, **33**(11): 130−136. DOI: 10.3969/j. issn. 1674-8425 (z).2019.11.019.

[17] DAI J H, WANG W, MI J S. Uncertainty Measurement for Interval-valued Information Systems[J]. *Inf Sci*, 2013, **251**: 63−78. DOI: 10.1016/j.ins.2013.06.047.