文章编号: 1000-6788(2009)06-0157-09

基于粗糙集的区间型数据离散化算法

谭 旭, 唐云岚, 陈英武

(国防科学技术大学 信息系统与管理学院, 长沙 410073)

摘 要 针对条件属性取值为区间型数据的离散化问题, 提出了一种新的基于粗糙集理论的离散化算法. 首先将粗糙集理论中上、下近似的概念进行扩展, 用以描述区间数对象间的距离和相似关系, 并通过定义相似度阈值来确定对象间的相似关系. 为了达到用最少的离散划分区间得到较好的离散化结果, 并合理地确定相似度阈值, 文章给出了粗糙熵的概念. 通过离散化属性的上、下近似粗糙熵值的计算以及该属性下各区间数对象的相似度矩阵的确定, 可以得到该属性下最终的离散化结果. 最后给出了一个烟叶质量等级评价的实例, 实验结果表明该算法是有效的.

关键词 粗糙集; 区间型数据; 离散化; 相似度矩阵; 相似度阈值; 粗糙熵

中图分类号 TP18 **文献标志码** A

Rough sets based discretization algorithm for interval data processing

TAN Xu, TANG Yun-lan, CHEN Ying-wu

(College of Information Systems & Management, National University of Defense Technology, Changsha 410073, China)

Abstract Aiming at discretizating condition attributes with interval value, we proposed a new discretization algorithm based on rough set theory. First, we extend the upper and lower approximation concept from rough set theory, so we can describe the distances and similarity relations among objects in an interval. Then similarity relations among the objects are determined by defining similarity threshold. In order to use minimal discretization intervals while obtaining optimal discretization results and define reasonable similarity threshold, we propose a new concept: rough entropy. By calculating upper and lower approximate rough entropy and determing similarity matrix of these objects, we can obtain the final discretization result. At the end of this paper, a tobacco leaf quality assessment example is given to prove the feasibility of this algorithm.

Keywords rough set theory; interval data; discretization; similarity matrix; similarity threshold; rough entropy

1 引言

数据挖掘与知识发现是当前的研究热点问题以及迅速发展的重要研究课题, 许多研究成果已成功地应用到了各个工程领域. 然而在大部分的实际应用中, 我们常遇到的是非精确或模糊性的数据. 如何从非精确或模糊性的数据表中获取有用的规则知识, 是我们所面临的一个有挑战性的难题. 粗糙集理论作为一种处理不确定性数据的软计算方法, 自 20 世纪 80 年代由波兰科学家 Pawlak[1] 提出以来, 已获得越来越广泛的关注. 目前粗糙集方法已被成功地应用于知识提取、机器学习、决策分析、模式识别、故障诊断等领域. 然而粗糙集方法只能处理离散型的数据, 对于属性值为连续型的数据必须先将其进行离散化处理. 而且, 即使对于离

收稿日期: 2008-02-02

作者简介: 谭旭 (1981–), 男, 湖南株洲人, 博士研究生, 目前从事粗糙集理论和多属性智能决策方面的研究; 唐云岚, 男, 博士研究生; 陈英武, 男, 教授, 博士, 博士生导师.

散型的数据, 也要将离散数据值进行合并以得到更高抽象层次的离散值, 从而更好地获取规则知识.

由于连续型数据的离散化结果能极大地影响到数据表的分类效果和规则知识的获取质量, 当前越来越受到学者们的关注和深入探讨. 基于粗糙集理论的传统连续型数据的离散化问题目前有了较深入的研究, 也有了较为丰富的研究成果. 根据离散化算法是否考虑具体的属性取值, 可分为监督离散化方法和非监督离散化方法[2]; 根据离散时考虑单个属性还是全体属性, 可分为全局离散化方法和局部离散化方法[3]; 根据离散划分是分类前还是分类过程中进行, 可分为静态离散化方法和动态离散化方法[4]. 不论采用何种离散化方法, 目的为获取更好的离散化结果从而提高规则知识的分类精度. 针对区间数的处理, 常见的有: 将区间型数据近似成精确型或符号型数据[5−6], 聚类整合区间型数据[7−8], 定义区间数之间的相离度[9], 区间回归分析[10]. 然而, 对于数据表中的数据取值为非精确或模糊型的数据, 尤其是区间型的数据, 如何对其进行离散化处理并利用粗糙集方法对其进行规则知识提取, 目前还没有很多的研究成果, 仍处于探究阶段. Pawlak[11] 尝试把"区间数" 的概念引入数值集合的粗糙集约简, 文献 [12] 提出了一种灰度粗糙集模型, 并从理论上阐述了该模型处理区间型数据的优越性, Leung[13] 探讨了区间数下的粗糙集理论的规则获取, 通过定义对象间相互的错误分类率来进行对象间相容度的计算, 但其阈值完全靠人为的设定.

本文通过扩展粗糙集上、下近似的概念来建立条件属性下各区间数对象间的相似关系, 并且定义粗糙熵来自动调整相似度阈值来达到对区间数属性的最优离散化. 全文利用粗糙集方法来解决不确定性 (区间型) 数据的离散化问题, 并用于知识的挖掘和推理, 给出了一条全新的解决思路, 并为粗糙集理论的更大范围的应用做了有益的尝试.

2 离散化问题与粗糙集的基本描述

定义 1[14] 决策表 T 可以表达为一个有序四元组 $T=\{U, C\cup D, V, f\}$. $U=\{o_1, o_2, \cdots, o_n\}$ 为决策表中全体数据对象的集合, $C=\{c_1, c_2, \cdots, c_m\}$ 为条件属性集, 反映对象的特征; D 为决策属性集, 反映对象的类别. $V=\bigcup\limits_{a\in C\cup D} V_a$ 为属性值的集合, f 为信息函数, 用于确定 U 中每一个对象在各个属性下的取值.

本文中集合 V 中的取值可以是精确型的数值数据, 也可以是区间型的数值数据, 亦可以为符号型的描述数据. 通常假定决策属性上的取值为离散型数据, 本文仅讨论条件属性集上的离散化问题. 下面给出条件属性集和决策属性离散化划分的定义[14].

定义 2 设 T 为决策表, U 为该决策表的有限论域, 决策属性 D 在 U 上的划分为 $U/ind(D)=\{G_1, G_2, \cdots, G_s\}$. $c_l\in C$ $(l=1,2,\cdots,m)$ 为任意一个取值为连续型数据的条件属性, $U/ind\{a_l\}=\{H_1^l, H_2^l, \cdots, H_t^l\}$ 为 U 在该条件属性 c_l 上的一种离散化结果. 这样, 决策表 T 在决策属性 D 和条件属性 c_l 的划分下, 可以得到一个相关的划分矩阵. 矩阵元素 X_{mn}^l $(\mathrm{m}=1,2,\cdots,t;\ n=1,2,\cdots,s)$ 为同时属于 H_m^l 分类和 G_n 分类的所有对象集合, $\left|X_{mn}^l\right|$ 为这些对象的数目.

定义 3 设 $R\subseteq C$, 集合 $R_-(Z)=\{z\in U, [z]_R\subseteq Z\}$ 表示为 Z 的 R 下近似集, 即根据知识 R, U 中能完全确定地归入集合 Z 的对象的集合; 集合 $R^-(Z)=\{z\in U, [z]_R\cap Z\neq\phi\}$ 表示为 Z 的 R 上近似集, 即根据知识 R, U 中可能归入集合 Z 的对象的集合. 其中, $[z]_R$ 等价关系 R 下包含对象 z 的等价类. $\alpha_R(Z)=Card(R_-(Z))/Card(U)$ 表示根据 R 能正确分类的对象的比率.

这里的知识 R 可以理解为一种广义的知识, 在本文可以是相似度关系, 集合 Z 也可以理解为某种阈值约束. $\alpha_R(Z)$ 可以用来评价条件属性离散划分的质量.

定义 4 设 P 和 Q 为 U 上的等价关系, Q 的 P 正域可以表示为 $Pos_P(Q)=\cup P_-(Z)$ $(Z\in U/Q)$, 即 U 中所有使用分类 U/P 所表达的知识能正确划分到 U/Q 的等价类中的对象所构成的集合. 通常, U/Q 为决策表中决策属性所得的划分, U/P 为离散化后的条件属性集所给出的划分.

3 粗糙集下区间数的处理

在实际问题中, 很多数据都是非精确或模糊的, 然而这些数据通常都可以归结为区间型数据的表达方式,

即允许数值在某个取值范围 δ 内变化. 下面将讨论如何处理决策表中的区间型数据, 进而将这些数据进行离散化便于进行粗糙集的规则知识提取.

给定区间型数据决策表 T, 如定义 1. 条件属性集为 $\{c_1, c_2, \cdots, c_m\}$, 对象在各条件属性下的取值均为区间型连续数据; 决策属性为 d, 对象在该决策属性下为离散型的取值, 对象集合在该决策属性下的划分为 $\{G_1, G_2, \cdots, G_s\}$. 本文拟采用局部离散化方法, 即基于决策属性的划分下, 分别考虑各个条件属性下对象集合的离散化处理. 给定条件属性 c_l, 任取对象 o_i 和 $o_j (i, j = 1, 2, \cdots, n;\ i \neq j)$, 它们在属性 c_l 下的取值为非精确的数值, 即分别为区间数 $[a_i^l, b_i^l]$ 和 $[a_j^l, b_j^l]$(如果某对象在该属性下的取值为精确数, 那么该区间数的上下限为相等的数值). 记 $a_{\min}^l = \min\limits_{h=1,\cdots,n} a_h^l$, $b_{\max}^l = \max\limits_{h=1,\cdots,n} b_h^l$. 下面考虑对象 o_i 和 o_j 在属性 c_l 下的上、下近似的距离. 根据两区间数之间的位置关系, 可以分为六种情况予以分别讨论, 如图 1 所示.

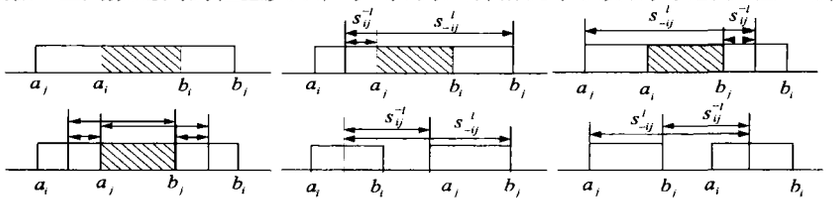

图 1　区间数对象间的位置关系

定义 5　对象 o_i 与 o_j 在属性 c_l 下的上近似的距离可以表示成 s_{ij}^{-l}

$$s_{ij}^{-l} = \begin{cases} 0, & a_j \le a_i < b_i \le b_j \\ p_{a_i \to a_j}\left(a_j - \dfrac{a_i + a_j}{2}\right), & a_i < a_j < b_i < b_j \\ p_{b_j \to b_i}\left(\dfrac{b_i + b_j}{2} - b_j\right), & a_j < a_i < b_j < b_i \\ p_{a_i \to a_j}\left(a_j - \dfrac{a_i + a_j}{2}\right) + p_{b_j \to b_i}\left(\dfrac{b_i + b_j}{2} - b_j\right), & a_i < a_j < b_j < b_i \\ a_j - \dfrac{a_i + b_i}{2}, & a_i < b_i < a_j < b_j \\ \dfrac{a_i + b_i}{2} - b_j, & a_j < b_j < a_i < b_i \end{cases}$$

其中 p 为概率, $p_{a_i \to a_j}$ 表示属性值落在区间 $[a_i, a_j]$ 内的概率, 这里假定数值的概率分布都为均匀分布. s_{ij}^{-l} 是区间数 $[a_i^l, b_i^l]$ 与区间数 $[a_j^l, b_j^l]$ 完全相似所需要的最小距离.

定义 6　对象 o_i 与 o_j 在属性 c_l 下的下近似的距离可以表示成 s_{-ij}^{l}

$$s_{-ij}^{l} = \begin{cases} 0, & a_j \le a_i < b_i \le b_j \\ p_{a_i \to a_j}\left(b_j - \dfrac{a_i + a_j}{2}\right), & a_i < a_j < b_i < b_j \\ p_{a_i \to a_j}\left(\dfrac{b_i + b_j}{2} - a_j\right), & a_j < a_i < b_j < b_i \\ p_{a_i \to a_j}\left(b_j - \dfrac{a_i + a_j}{2}\right) + p_{b_j \to b_i}\left(\dfrac{b_i + b_j}{2} - a_j\right), & a_i < a_j < b_j < b_i \\ b_j - \dfrac{a_i + b_i}{2}, & a_i < b_i < a_j < b_j \\ \dfrac{a_i + b_i}{2} - a_j, & a_j < b_j < a_i < b_i \end{cases}$$

s_{-ij}^{l} 是区间数 $[a_i^l, b_i^l]$ 与区间数 $[a_j^l, b_j^l]$ 完全相似所需要的最大距离. 通常, $s_{-ij}^{l} \neq s_{-ji}^{l}, s_{ij}^{-l} \neq s_{ji}^{-l}$, 即任意两对象在某属性下的区间数取值间的距离不具有对称性.

记 $o_{-ij}^{l} = \max\{s_{-ij}^{l}, s_{-ji}^{l}\}$, $o_{ij}^{-l} = \max\{s_{ij}^{-l}, s_{ji}^{-l}\}$, 则 $o_{-ij}^{l} = o_{-ji}^{l}$, $o_{ij}^{-l} = o_{ji}^{-l}$.

定义 7 $clo^l_{-ij}=(b^l_{\max}-a^l_{\min}-o^l_{-ij})/(b^l_{\max}-a^l_{\min})$ 为对象 o_i 与对象 o_j 在属性 c_l 下的下近似相似度; $clo^{-l}_{ij}=(b^l_{\max}-a^l_{\min}-o^{-l}_{ij})/(b^l_{\max}-a^l_{\min})$ 为对象 o_i 与对象 o_j 在属性 c_l 下的上近似相似度.

clo^l_{-ij} 和 clo^{-l}_{ij} 均在 $[0,1]$ 内取值, $clo^l_{-ij}\to 0$ 说明在下近似范围内, 两取值区间越相异; $clo^l_{-ij}\to 1$ 说明在下近似范围内, 两取值区间越接近. 并且区间数之间的距离越大, 则两区间之间的下近似相似度越低. 上近似相似度取值分析同理.

在给出了区间数之间的上、下近似相似度后, 可以分别构造出决策表 T 中对象集合在区间数的条件属性 c_l 下的 $n\times n$ 上、下近似相似矩阵 M^l_- 和 M^{-l}, 其中 $m^l_{-ij}=clo^l_{-ij}$, $m^{-l}_{ij}=clo^{-l}_{ij}$. 显然 $m^l_{-ij}=m^l_{-ji}$, $m^{-l}_{ij}=m^{-l}_{ji}$ 为对称矩阵.

在得到上、下近似相似矩阵 M^l_- 和 M^{-l} 后, 根据设定的相似度阈值 η 来确定对象集合在条件属性 c_l 下的离散划分. 如果 $m^l_{-ij}\geq\eta$, 则对象 o_i 与 o_j 在阈值 η 下的下近似是属于同一离散划分的, 即在阈值 η 下, 对象 o_i 与 o_j 是 "一定" 属于同一离散划分. 如果 $m^{-l}_{ij}\geq\eta$, 则对象 o_i 与 o_j 在阈值 η 下的上近似是属于同一离散划分的, 即在阈值 η 下, 对象 o_i 与 o_j 是 "可能" 属于同一离散划分的. 如图 2 所示.

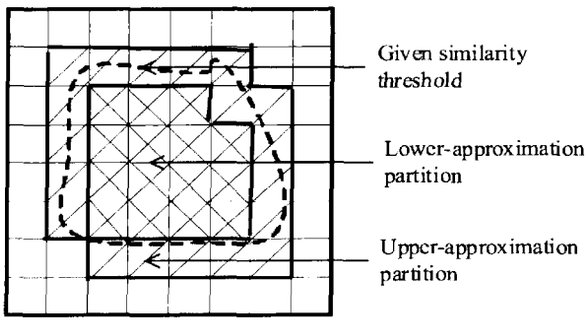

图 2 上、下近似关系下的对象划分

例 1 给定区间型数据的决策表中各对象在条件属性 c_l 上的取值 (如表 1 示).

表 1 给定某属性下的区间型数据的处理示例

属性	对象				
	o_1	o_2	o_3	o_4	o_5
c_l	[1.7, 1.9]	[1.8, 2.2]	[3.1, 3.7]	[4.9, 5.6]	[4.8, 6.1]

容易得知, $a_{\min}=1.7$, $b_{\max}=6.1$

可以得到上、下近似距离矩阵:

$$O^l_-=\begin{bmatrix}0\\0.26&0\\1.90&1.70&0\\3.80&3.60&2.20&0\\4.30&4.10&2.70&0.42&0\end{bmatrix}\quad O^{-l}=\begin{bmatrix}0\\0.11&0\\1.50&1.20&0\\3.35&3.05&1.55&0\\3.55&3.25&1.75&0.10&0\end{bmatrix}$$

相应的上、下近似相似矩阵为:

$$M^l_-=\begin{bmatrix}1\\0.94&1\\0.57&0.61&1\\0.14&0.18&0.50&1\\0.02&0.07&0.39&0.90&1\end{bmatrix}\quad M^{-l}=\begin{bmatrix}1\\0.97&1\\0.70&0.73&1\\0.24&0.31&0.64&1\\0.19&0.26&0.60&0.98&1\end{bmatrix}$$

给定 $\eta=0.7$, $\{o_1,o_2\}$, $\{o_3\}$, $\{o_4,o_5\}$ 为属性 c_l 下的下近似离散划分, 相应的离散化划分区间为 $[1.7,2.2]$, $[3.1,3.7]$, $[4.8,6.1]$. $\{o_1,o_2,o_3\}$ 和 $\{o_4,o_5\}$ 为在属性 c_l 下的上近似离散划分, 相应的离散化划分区间为 $[1.7,3.7]$, $[4.8,6.1]$.

这里, 相似度阈值 η 是主观给出的, 如何合理地确定 η 值, 选择怎样的离散划分作为属性下的最终离散划分? 以及如何去进行粗糙规则提取, 在下一节我们将给出详细的算法.

4 基于粗糙集的区间型数据离散化算法

在连续型精确数据的离散化中, 基于信息熵的离散化方法是当前普遍采用并行之有效的一种方法[15]. 该方法的目的是使得在某个条件属性的离散化划分区间中, 其所有区间内的对象集合越 "纯" 越好. 即对象集合在决策属性的分类之后, 某条件属性中所有划分区间内的对象集合属于越少的决策划分类别, 则该离散化划分越优越. 但这样的划分会趋向于将条件属性下的离散化划分粒度推向 "越细越好", 一个极端的情况是条件属性下的每个取值都作为一个断点划分, 这样的信息熵一定最小. 显然, 这样的离散化并不是我们理想的结果. 在条件属性的离散化中, 我们需要寻找一种合适的划分 "粒度", "粒度" 越细, 则得到的推理规则越冗余; "粒度" 越粗, 则容易导致不一致性. 一种理想的结果是, 在获得尽可能少的离散化划分区间的同时达到较优的划分结果. 下面我们提出一种 "粗糙熵" 的方法来调节相似度阈值 η, 从而达到对连续区间型数据的离散化.

设决策表中的对象集合在决策属性上的划分为 $\{G_1, G_2, \cdots, G_s\}$, $|G_s|$ 为划分到类 G_s 的对象数目. 考虑连续区间型条件属性 c_l 下的离散化. 假设在相似度阈值 η 下, 根据第 3 节所提出的区间数处理方法, 条件属性 c_l 下近似下的离散划分为 $\{H^l_{-1}, H^l_{-2}, \cdots, H^l_{-u}\}$, $\left|H^l_{-u}\right|$ 为划分到类 H^l_{-u} 的对象数目; 上近似下的离散划分为 $\{H^{l-}_1, H^{l-}_2, \cdots, H^{l-}_v\}$, $\left|H^{l-}_v\right|$ 为划分到类 H^{l-}_v 的对象数目. 下近似下, 划分于类 H^l_{-u} 的对象属于决策划分 G_s 的对象个数记为 $\left|N_{H^l_{-u}\to G_s}\right|$, 上近似下, 划分于类 H^{l-}_v 的对象属于决策划分 G_s 的对象个数记为 $\left|N_{H^{l-}_v\to G_s}\right|$.

定义 8 条件属性 c_l 在下近似划分 $H^l_{-r}(r=1,2,\cdots,u)$ 内的下近似信息熵为 $I_-(c^l_r) = -\sum_{k=1}^{s} pro_k \ln pro_k$, $pro_k = \left|N_{H^l_{-u}\to G_s}\right| \Big/ \left|H^l_{-r}\right|$ 表示该划分内的对象属于决策类 G_k 的概率, 则属性 c_l 在离散划分后的下近似粗糙熵可以定义为 $I_-(c_l) = \alpha \sum_{r=1}^{u} \left(\left|H^l_{-r}\right|/n\right) I_-(c^l_r)$, 其中 α 为粗糙因子, $\alpha = \sum_{r=1}^{u} (\lambda_r/s)$. λ_r 表示对象集在划分 H^l_{-r} 内取值所属决策类的数目, s 为决策类的总数目.

定义 9 某条件属性 c_l 在上近似划分 $H^{l-}_q(q=1,2,\cdots,v)$ 内的上近似信息熵为 $I^-(c^l_q) = -\sum_{k=1}^{s} pro_k \ln pro_k$, $pro_k = \left|N_{H^{l-}_q\to G_k}\right| \Big/ \left|H^{l-}_q\right|$ 表示该划分内的对象属于决策类 G_k 的概率, 则属性 c_l 在离散划分后的上近似粗糙熵可以定义为 $I^-(c_l) = \alpha \sum_{q=1}^{v} \left(\left|H^{l-}_q\right|/n\right) I^-(c^l_q)$, $\alpha = \sum_{q=1}^{v} (\lambda_q/s)$. λ_q 表示对象集在划分 H^{l-}_q 内取值所属决策类的数目, s 为决策类的总数目.

通过粗糙熵的定义, 可以依据 "在条件属性上获得尽可能少的离散划分区间的同时达到较优的划分结果" 来调整相似度阈值 η. 粗糙熵越小, 说明该离散划分得到的划分类 (区间) 越少, 同时表明在某个划分内的 "混乱" 程度越小, 即某个划分内的对象属于尽可能少的决策类别, 这样的离散划分是较优的划分.

例 2 决策表仍参照例 1 中的数据, 假设该决策表中的对象被决策属性划分为决策类 $\{o_1, o_2, o_3, o_4\}$ 和 $\{o_5\}$. 在相似度阈值 $\eta = 0.7$ 下的条件属性离散划分见例 1 中的结果.

则, 下近似下的粗糙熵为:

$$I_-(c_l) = \alpha \sum_{r=1}^{3} \left(\left|H^l_{-r}\right|/n\right) I_-(c^l_r) = 2^* \left(2/5^*0 + 1/5^*0 + 2/5\left(-\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2}\right)\right)$$
$$= 2^*(0+0+0.2773) = 0.5546$$

上近似下的粗糙熵为:

$$I^-(c_l) = \alpha \sum_{q=1}^{2} \left(\left|H^{l-}_q\right|/n\right) I^-(c^l_q) = 3/2^* \left(3/5 * 0 + 2/5^* \left(-\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2}\right)\right)$$
$$= 3/2^*(0+0.2773) = 0.4159$$

可以看到, 本例中在给定相似度阈值下, 上近似粗糙熵要小于下近似粗糙熵, 是因为在划分效果等同的情况下, 划分类别较少的离散化划分将更加体现其优越性. 而原有的信息熵将无法区分这两种划分, 都等同于 0.2773.

定理 1 对于决策表 T 中某个区间型条件属性 c_l, 设 $I^-(c_l)$ 为该属性下的上近似粗糙熵, $I_-(c_l)$ 为该属性下的下近似粗糙熵, 如果上近似粗糙熵 $I^-(c_l)$ 在下近似粗糙熵 $I_-(c_l)$ 取最小值的情况下, 取值最小, 那么这样的离散划分为粗糙熵意义下的最优划分.

证明 设 $I_{-\min}$ 为该区间型属性离散化划分的最小下近似粗糙熵值, $I^-_{\min}$ 为下近似粗糙熵值取最小值的情况下, 上近似粗糙熵能够取得的最小值. 根据粗糙熵的定义知道, 下近似粗糙熵是基于下近似相似下的划分, 即下近似粗糙熵代表的是 "确定" 的划分而得到的熵值. 对应的, 上近似粗糙熵代表的是 "可能" 的划分而得到的熵值. 显然, 下近似粗糙熵的重要度大于上近似粗糙熵. $\forall I > I_{-\min}$, 相比 $I_{-\min}$, 在下近似粗糙熵值取值为 I_- 的条件下, 根据定义 8 中的粗糙熵定义, 这样的离散划分在下近似相似度下, 使得更多的决策属性类充斥于各个离散划分区间, 不确定因素将变大, 并且划分区间数目也更多, 显然这样的离散化划分显然不会比取值为 $I_{-\min}$ 时更优. $\forall I_- = I_{-\min}$ 且 $I^- > I^-_{\min}$, 即在下近似下的粗糙熵取到最小而上近似粗糙熵大于最小值, 同理, 这样的离散划分相比 $I_- = I_{-\min}$ 且 $I^- = I^-_{\min}$ 的条件下为不占优划分.

这样, 通过计算上、下粗糙熵的大小, 我们可以来确定离散划分的相似度阈值 η. 具体的离散化算法描述如下:

Step 1 在决策表中依次选取区间型连续条件属性 $c_l \in C$ $(l = 1, 2, \cdots, m)$, 设定初始相似度阈值 $\eta_0 \in [0.5, 1)$, 设置相似度调整步长 $\xi = 0.01$, 设定初始上近似粗糙熵 I^-_0, 下近似粗糙熵 I_{-0};

Step 2 计算该属性下取值为区间数的对象集的上、下近似距离矩阵 O^{-l}、O^l_-, 并转换成相似矩阵 M^{-l}、M^l_-;

Step 3 得到在相似度阈值 η_0 下, 该属性下对象集合的上、下近似离散划分 H^{l-}、H^l_-;

Step 4 计算该离散划分下的下近似粗糙熵 I_- 和上近似粗糙熵 I^-, if $I_- < I_{-0}$, then $I_{-0} := I_-$, $\eta := \eta_0$, $\eta_0 = \eta_0 + \xi$; if $I_- = I_{-0}$ and $I^- < I^-_0$, then $I^-_0 := I^-$, $\eta := \eta_0$, $\eta_0 = \eta_0 + \xi$; else $\eta_0 = \eta_0 + \xi$;

Step 5 if $\eta_0 \notin [0.5, 1)$, 得到该属性下的最终 η 值, 选取在 η 下上、下粗糙熵值较小的划分作为该属性下的最终划分结果, else 转 Step3;

Step 6 if 该决策表中所有连续区间型属性全部离散化完毕, 结束; else 转 Step1.

将所有带有连续区间型数据的属性进行如上离散化后, 再进行属性的约简, 可以得到 if$\cdots$ then$\cdots$ 形式的规则集, 也即挖掘得到该决策表中的知识. 这些规则有些是确定型的规则, 有些是带可信度的规则. 根据这些规则便可以进行推理.

5 实例

以烤烟烟叶的品质分类为例来进一步阐释本文的算法. 根据烤烟的 42 等级国标分类分级方法, 为了便于从宏观上控制配方等级的使用, 我们又将这 42 等级的烟叶按质量分为上等烟、中等烟和下等烟[16]. 目前, 烤烟烟叶品质的分类仍然没有一个完全统一的认识, 尚属于一个宽泛而模糊的概念, 根据不同的卷烟厂有着不同的定义和理解. 究其原因, 还是因为没有一个精准而合理的分类方法, 主要还是以传统的 "眼观手摸" 的方式主导. 然而, 烟叶的化学成分是烟叶中比较稳定的因素, 能直接反映烟叶的质量好坏[17]. 我们尝试根据烟叶的化学成分来进行烟叶品质的探索分析. 本文采集了 30 条某卷烟厂烤烟烟叶的化学成分及其对应的品质等级数据. 由于同一等级的烟叶来自不同的地域, 所以化学成分存在某种程度的偏差; 另外, 即使同一批次的烟叶, 也存在化学成分的测量误差. 本文将这些数据予以了详细的记录 (见表 2). 其中, 22 条数据用作规则提取, 另外 8 条数据作为测试.

根据本文提出的离散化算法, 参照决策属性 "烟叶等级", 逐一对所有区间型条件属性进行离散化. 这里, 设定相似度阈值在 $[0.6, 1]$ 内取值, 调整步长设定为 $\xi = 0.01$. 以 "总挥发碱" 为例, 通过计算上、下近似距离矩阵和相似矩阵, 得到在该属性上所有对象之间取值的相似度. 在设定的步长下, 通过搜索计算在不同相似度阈值取值的条件下, 上、下近似粗糙熵的取值变化, 得到最优的相似度阈值. 通过图 3 可以看到, 在相似度阈值取值为 0.64 时, 下近似粗糙熵值为全局最小 (2.245), 并且此刻的上近似粗糙熵也是在下近似粗糙熵值取值全局最小时所能够取到的最小值 (2.338). 故, 确定对属性 "总挥发碱" 离散化的相似度阈值为 0.64. 在

表 2　某卷烟厂烤烟烟叶化学成分数据品质等级

对象	烤烟化学成分指标										烟叶等级
	总挥发碱	总氮	烟碱	还原糖	总酸	总挥发酸	全纤维素	总灰分	钾	氯	
1	[0.3, 0.35]	[1.9, 2.1]	[2.3, 2.34]	[16, 18]	[14, 14.5]	[0.57, 0.58]	[16.9, 17.1]	[12.2, 12.3]	[1.5, 1.52]	[0.44, 0.46]	上
2	[0.3, 0.35]	[1.85, 1.9]	[2.1, 2.14]	[16, 18]	[14, 14.5]	[0.57, 0.58]	[16.1, 16.2]	[13.7, 13.8]	[1.3, 1.33]	[0.31, 0.33]	上
3	[0.26, 0.27]	[1.9, 2.1]	[2.4, 2.6]	[18.8, 19]	[13.5, 13.7]	[0.57, 0.58]	[19.1, 19.3]	[14, 14.1]	[1.34, 1.36]	[0.51, 0.53]	上
4	[0.37, 0.38]	[2.26, 2.3]	[1.86, 1.9]	[16, 18]	[14.8, 14.9]	[0.75, 0.9]	[12, 15]	[17.2, 17.3]	[1.34, 1.37]	[0.44, 0.46]	上
5	[0.28, 0.29]	[1.68, 1.72]	[2.86, 2.92]	[16.8, 17.0]	[14, 14.5]	[0.49, 0.5]	[17.3, 17.6]	[15.6, 15.7]	[1.49, 1.51]	[0.2, 0.3]	上
6	[0.4, 0.42]	[2.22, 2.26]	[2.19, 2.22]	[18.9, 19.0]	[14.7, 14.8]	[0.75, 0.85]	[16, 16.2]	[18, 18.1]	[1.32, 1.34]	[0.43, 0.45]	上
7	[0.27, 0.28]	[1.74, 1.78]	[2.83, 2.88]	[12.2, 12.3]	[12.2, 12.3]	[0.47, 0.48]	[18.9, 19.1]	[16.2, 16.3]	[1.31, 1.34]	[0.74, 0.76]	中
8	[0.23, 0.24]	[1.74, 1.78]	[3.21, 3.23]	[11.2, 11.4]	[11.1, 11.2]	[0.44, 0.45]	[21.2, 21.4]	[17.2, 17.3]	[1.07, 1.09]	[0.44, 0.45]	中
9	[0.19, 0.2]	[1.14, 1.18]	[3, 3.05]	[15.8, 16]	[11, 11.1]	[0.41, 0.42]	[19.8, 20]	[17.7, 17.8]	[1.53, 1.56]	[0.48, 0.5]	中
10	[0.22, 0.23]	[1.3, 1.36]	[3.02, 3.07]	[15, 15.2]	[11.2, 11.3]	[0.4, 0.41]	[21.4, 21.6]	[15.5, 15.6]	[1.49, 1.50]	[0.44, 0.46]	中
11	[0.27, 0.28]	[1.8, 1.83]	[1.82, 1.86]	[18.3, 18.5]	[13, 13.1]	[0.47, 0.48]	[21.1, 21.3]	[13.2, 13.3]	[1.53, 1.56]	[0.79, 0.81]	中
12	[0.3, 0.35]	[2.14, 2.16]	[2.36, 2.39]	[19.5, 19.7]	[15.5, 15.6]	[0.69, 0.7]	[18.3, 18.5]	[18.2, 18.3]	[1.21, 1.23]	[0.39, 0.4]	中
13	[0.18, 0.19]	[1.9, 2.0]	[2.32, 2.36]	[18.9, 19.2]	[13.3, 13.5]	[0.52, 0.53]	[21.3, 21.5]	[14, 14.2]	[1.35, 1.36]	[0.79, 0.81]	中
14	[0.47, 0.48]	[2.14, 2.16]	[2.4, 2.6]	[21, 21.3]	[15, 15.1]	[0.75, 0.9]	[20.2, 20.4]	[18.1, 18.2]	[1.27, 1.29]	[0.41, 0.43]	中
15	[0.39, 0.4]	[2.46, 2.5]	[2.06, 2.1]	[20.3, 20.6]	[15.4, 15.5]	[0.64, 0.65]	[19.1, 19.3]	[17.4, 17.5]	[1.24, 1.26]	[0.57, 0.59]	中
16	[0.18, 0.2]	[1.19, 1.22]	[1.73, 1.76]	[16, 18]	[11.7, 11.8]	[0.65, 0.66]	[24.1, 24.3]	[17.2, 17.3]	[1.36, 1.38]	[0.44, 0.45]	中
17	[0.22, 0.23]	[1.48, 1.5]	[1.70, 1.75]	[11.4, 11.8]	[12.7, 12.8]	[0.43, 0.44]	[22.3, 22.5]	[13.4, 13.5]	[1.09, 1.11]	[1.07, 1.09]	下
18	[0.51, 0.52]	[2.38, 2.42]	[1.59, 1.61]	[22.8, 23]	[17.7, 17.8]	[0.67, 0.68]	[18.2, 18.4]	[17.2, 17.3]	[1.34, 1.36]	[0.92, 0.94]	下
19	[0.22, 0.23]	[1.43, 1.47]	[3.08, 3.12]	[14.1, 14.3]	[11.9, 12]	[0.37, 0.38]	[22.2, 22.4]	[16.4, 16.6]	[1.34, 1.36]	[0.8, 0.82]	下
20	[0.2, 0.21]	[1.38, 1.42]	[1.54, 1.56]	[19.7, 19.9]	[12.7, 12.8]	[0.6, 0.61]	0	[16.8, 17]	[1.42, 1.44]	[0.44, 0.45]	下
21	[0.22, 0.23]	[1.27, 1.31]	[3.01, 3.05]	[19.9, 20.2]	[13.2, 13.3]	[0.6, 0.61]	0	[17.9, 18.2]	[1.16, 1.18]	[0.67, 0.69]	下
22	[0.51, 0.52]	[2.72, 2.75]	[1.74, 1.78]	[20.2, 20.4]	[15.9, 16]	[0.67, 0.68]	[19.3, 19.6]	[15.6, 15.7]	[1.15, 1.17]	[0.5, 0.52]	下

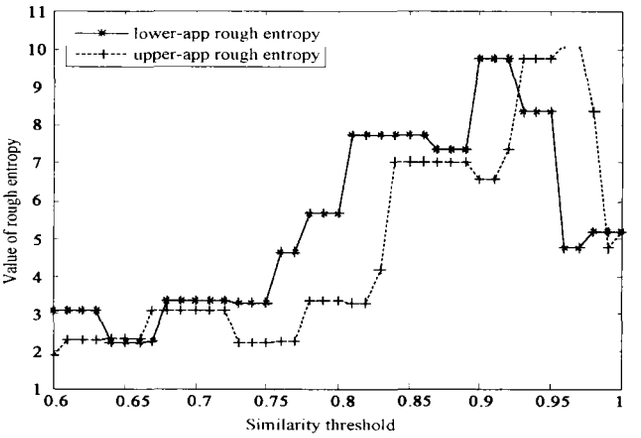

图 3　属性 "总挥发碱" 的离散化寻优

给定相似度阈值 0.64, 下近似粗糙熵小于上近似粗糙熵. 最终离散划分结果为 {1, 2, 3, 4, 5, 6, 7, 12, 15}, {8, 9, 10, 11, 13, 16, 17, 19, 20, 21}, {14, 18, 22}. 剩余 9 个条件属性的离散化结果见表 3, 粗体的数据表示为该属性下最终选择的离散化结果.

将决策表中的数据按本文的算法进行离散化后, 再进行属性的约简和属性值的约简. 根据属性重要度的计算, 确定约简后的属性集为{总挥发碱, 总氮, 烟碱, 还原糖, 总灰分}. 最终, 我们得到如下 6 条推理规则:

Rule1: if 还原糖 $\in [15, 19.2]$ and 烟碱 $\in [1.82, 2.92]$ and 总挥发碱 $\in [0.28, 0.42]$ then 上等烟

Rule2: if 还原糖 $\in [15, 19.2]$ and 总氮 $\in [1.68, 2.3]$ and 总挥发碱 $\in [0.28, 0.42]$ then 上等烟

Rule3: if 还原糖 $\in [11.2, 14.3]$ and 烟碱 $\in [1.82, 2.92]$ then 中等烟

Rule4: if 还原糖 $\in [11.2, 14.3]$ and 烟碱 $\in [3.01, 3.23]$ and 总氮 $\in [1.68, 2.3]$ then 中等烟

Rule5: if 还原糖 $\in [15, 19.2]$ and 总挥发碱 $\in [0.18, 0.24]$ then 中等烟

Rule6: if 总灰分 $\in [15.5, 18.3]$ and (还原糖 $\in [11.2, 14.3]$ or 还原糖 $\in [19.5, 20.6]$) and (烟碱 $\in [1.54, 1.78]$ or 烟碱 $\in [3.01, 3.23]$) then 下等烟

表 3 各区间型数据条件属生下的离散化结果

属性		离散划分结果			粗糙熵	阈值
总挥发碱	下近似	**{1, 2, 3, 4, 5, 6, 7, 12, 15}**	**{8, 9, 10, 11, 13, 16, 17, 19, 20, 21}**	**{14, 18, 22}**	2.245	0.64
	上近似	{1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 15, 17, 19, 20, 21}	{9, 13, 16}	{14, 18, 22}	2.338	
总氮	下近似	**{1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14}**	**{9, 10, 16, 17, 19, 20, 21}**	**{15, 18, 22}**	2.181	0.68
	上近似	{1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 18, 22}	{9, 10, 16, 17, 19, 20, 21}		2.293	
烟碱	下近似	{1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 17, 21, 22}	{8, 9, 10, 19}	{18, 20}	4.621	0.69
	上近似	**{1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15}**	**{8, 9, 10, 19, 21}**	**{16, 17, 18, 20, 22}**	3.574	
还原糖	下近似	{1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 16, 19}	{7, 8, 17}	{14, 15, 18, 20, 21, 22}	2.587	0.74
	上近似	**{1, 2, 3, 4, 5, 6, 9, 10, 11, 13, 16}**	**{7, 8, 17, 19}**	**{12, 14, 15, 18, 20, 21, 22}**	2.011	
总酸	下近似	{1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 17, 19, 20, 21, 22}	{8, 9, 10, 16}	{18}	1.822	0.67
	上近似	**{1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}**	**{8, 9, 10}**		1.816	
总挥发酸	下近似	**{1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22}**	**{4, 6, 14}**		1.888	0.70
	上近似	{1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22}	{4, 6, 14}		1.888	
全纤维素	下近似	{1, 2, 3, 4, 5, 6, 7, 9, 12, 14, 15, 18, 22}	{8, 10, 11, 13, 16, 17, 19}	{20, 21}	1.877	0.70
	上近似	**{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 22}**	**{20, 21}**		1.373	
总灰分	下近似	**{1, 2, 3, 11, 13, 17}**	**{4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 18, 19, 20, 21, 22}**		3.512	0.68
	上近似	{1, 2, 3, 11, 13, 17}	{4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 18, 19, 20, 21, 22}		3.512	
钾	下近似	{1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 16, 18, 19, 20}	{12, 14, 15, 21, 22}	{8, 17}	4.073	0.60
	上近似	**{1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 16, 18, 19, 20}**	**{8, 12, 14, 15, 17, 21, 22}**		2.045	
氯	下近似	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 20, 21, 22}	{7, 11, 13}	{17, 18, 19}	1.969	0.66
	上近似	**{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 21, 22}**	**{17, 18, 19}**		1.324	

根据国际普遍认同的烤烟烟叶质量评价参考[16], 还原糖对烟叶的吸味品质占有极其重要的影响, 它应与烟碱 (尼古丁) 和总氮有适当的比例. 其最佳值应在 15%左右. 而烟碱 (尼古丁) 与还原糖的最佳比例为 1:(6−8), 烟碱 (尼古丁) 与总氮的合适比例为 1:1. 总氮含量高的烟叶劲头和刺激大, 过低则吃味差. 最佳值为 2.2%左右. 总灰份反应了烟叶的燃烧性, 低者燃烧性好. 从得到的推理规则集可以看到, 我们的推理规则是与这些评价标准相吻合的, 也从一个侧面反应出本离散化算法的合理性和正确性. 利用剩余的测试数据进行测试, 预测正确识别率为 75%, 拒识率为 12.5%, 误识率为 12.5%.

6 结语

粗糙集方法优点显著, 不需要任何辅助信息可以从决策表中很好地获取规则知识. 但是粗糙集方法只能处理离散型数据的固有缺陷很大程度地限制了其发展, 尤其是对于区间型数据的决策表, 利用粗糙集理论进行规则提取是尤为困难的. 然而在现实生活应用中, 由于种种的不确定性因素, 我们更常遇到的是以模糊、不确定、区间数的形式出现的数据. 本文尝试利用粗糙集上、下近似的观点来处理区间型数据的离散化问题, 并提出粗糙熵的思想来对数据进行离散化寻优计算. 最后将本算法应用到烤烟烟叶的品质评价分类中, 得到了令人鼓舞的结果. 使得烟叶品质分类摆脱了传统的人工模式, 实现了按照烟叶化学成分进行分类的自动化和精确化, 提高了品质分类的可信度. 同时, 也证实了本算法的普适性. 文章通过对决策表中含有模糊不确定型数据的离散化进行的有益尝试, 进一步扩大了粗糙集理论的适用范围.

参考文献

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

[2] Huan L, Farhad H, Manoranjan D. Discretization: An enabling technique[J]. Data Mining and Knowledge Discovery, 2002, 6: 393–423.

[3] Chmielewski M R, Grzymala-Busse J W. Global discretization of continuous attributes as preprocessing for machine learning[J]. International Journal of Approximate Reasoning, 1996, 15: 319– 331.

[4] 刘文军. 基于粗糙集的数据挖掘算法研究 [D]. 北京: 北京师范大学, 2004.
Liu W J. Research on data mining algorithms based on rough sets [D]. Beijing: Beijing Normal University, 2004.

[5] 谭旭, 高妍方, 陈英武. 区间型多数属性决策求解新方法 [J]. 系统工程与电子技术, 2007, 29(7): 1082–1085.
Tan X, Gao Y F, Chen Y W. New method for solving interval multi-attribute decision-making problem[J]. Systems Engineering and Electronics, 2007, 29(7): 1082–1085.

[6] Bock H H, Diday E. Anal Symbolic Data[M]. Berlin: Springer-Verlag, 2000.

[7] de Souza R M C R, de Carvalho F A T. Clustering of interval data based on city-block distances[J]. Pattern Recognition Letter, 2004, 25(3): 353–365.

[8] Asharaf S, Murty M N, Shevade S K. Rough set based incremental clustering of interval data[J]. Pattern Recognition Letter, 2006, 27: 515–519.

[9] 徐泽水. 不确定多属性决策方法及应用 [M]. 北京: 清华大学出版社, 2004.
Xu Z S. Uncertain Multiple Attribute Decision Making: Method and Applications[M]. Beijing: Tsinghua University Press, 2004.

[10] Tanaka H, Guo P. Possibilistic Data Analysis for Operations Research[M]. Heidelberg: Physica-Verlag, 1999.

[11] Pawlak Z, Skowron A. Rough sets and Boolean reasoning [J]. Information Sciences, 2007, 177: 41–73.

[12] Yamaguchi D, Li G, Nagai M. A grey-based rough approximation model for interval data processing[J]. Information Sciences, 2007, 177(21): 4727–4744.

[13] Leung Y, Fischer M M, Wu W Z, et al. A rough set approach for the discovery of classification rules in interval-valued information systems[J]. International J of Approximate Reasoning, 2008, 47(2): 233–246.

[14] 张文修, 仇国芳. 基于粗糙集的不确定决策 [M]. 北京: 清华大学出版社, 2005.
Zhang W X, Qiu G F. Uncertain Decision Making Based on Rough Sets[M]. Beijing: Tsinghua University Press, 2005.

[15] 谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法 [J]. 计算机学报, 2005, 28(9): 1570–1574.
Xie H, Cheng H Z, Niu D X. Discretization of continuous attributes in rough set theory based on information entropy[J]. Chinese Journal of Computers, 2005, 28(9): 1570–1574.

[16] 胡开文. 烟叶打叶复烤工艺与设备 [M]. 北京: 化学工业出版社, 2002.
Hu K W. Technique and Equipment in Tobacco Threshing and Redrying [M]. Beijing: Chemical Industry Press, 2002.

[17] 毛多斌, 马宇平, 梅业安. 卷烟配方和香精香料 [M]. 北京: 化学工业出版社, 2001.
Mao D B, Ma Y P, Mei Y A. Cigarette Blends with Flavor and Fragrance[M]. Beijing: Chemical Industry Press, 2001.