

引用格式:陈曦. 基于直觉模糊粒化的信息熵属性约简算法[J]. 微电子学与计算机,2020,37(1):38-45. [CHEN X. Attribute reduction algorithm based on the information entropy of intuitionistic fuzzy granulation[J]. Microelectronics & Computer,2020, 37(1):38-45.]

基于直觉模糊粒化的信息熵属性约简算法

陈 曦^{1,2}

(1 华中科技大学 管理学院, 湖北 武汉 430074;

2 武汉铁路职业技术学院, 湖北 武汉 430200)

摘 要:在直觉模糊关系中,对象之间通过隶属度和非隶属度的刻画使得拥有了更为优越的关系评估效果.为了对信息系统的确定性达到更好的度量,首先引入基于直觉模糊关系对信息系统进行直觉模糊粒化,然后在粒化的结果中依据隶属度和非隶属度分别定义了信息熵的概念,并将它们结合作为直觉模糊关系下信息系统的信息熵,最后根据该信息熵构造一种属性约简算法.实验结果表明提出的算法具有较优的属性约简性能.

关键词:粒计算;直觉模糊关系;隶属度;非隶属度;信息熵;属性约简

中图分类号: TP18

文献标识码: A

文章编号: 1000—7180(2020)01—0038—08

Attribute reduction algorithm based on the information entropy of intuitionistic fuzzy granulation

CHEN Xi^{1,2}

(1 School of Management, Huazhong University of Science and Technology, Wuhan 430074 ,China;

2 Wuhan Railway Vocational and Technical College, Wuhan 430200, China)

Abstract: In intuitionistic fuzzy relations, between objects have more superior relationship evaluation effect through the description of membership degree and non-membership degree. In order to achieve better measurement for information system. Firstly, the intuitionistic fuzzy granulation for information system based on intuitionistic fuzzy relations is introduced in this paper. And then, the concepts of entropy are respectively defined according to membership degree and non-membership degree in granulation results, in addition, combining them as the information entropy of the information system under the intuitionistic fuzzy relation. Finally, an attribute reduction algorithm is constructed according to the information entropy. Experimental results show that the proposed algorithm has better attribute reduction performance.

Key words: granular computing; intuitionistic fuzzy relationship; membership degree; non-membership degree; information entropy; attribute reduction

1 引言

粒计算理论^[1]是近年来提出的一种数据分析工具,它与粗糙集^[2]、模糊集^[3]等理论的结合,使得它在分析与处理不确定性数据方面表现出了强大的优

越性,目前已成功运用于数据挖掘^[4],机器学习^[5-6]等领域.

属性约简^[1-2]是粗糙集和粒计算中一项重要的应用,现实中的很多数据集中存在着大量的冗余属性,这些冗余的属性不仅增加分类训练的计算量,而

且还会降低分类性能^[5],属性约简的目的就是消除那些冗余的属性.目前学者们提出大量的属性约简方法,例如 Qian 等学者^[7]运用粗糙集中的等价关系粒化构造了属性约简算法,Dai 等学者^[8]和续欣莹等学者^[9]基于信息系统信息粒化的熵理论提出了相应的属性约简,Lin 等学者^[10]运用邻域粗糙集中邻域粒化构造了数值属性约简,Zhang 等学者^[11]和 Wang 等学者^[12]基于模糊粗糙集模型分别提出了高效的属性约简, Ma 等^[13]学者在决策粗糙集理论中提出了保持决策区域的近似约简.在增量式学习中, Jing 等^[14-17]学者提出了多种增量式属性约简方法,在并行计算领域中, Chen 等^[18]学者和 Zhang 等^[19]学者也分别提出了相应的并行属性约简算法,这些算法的不断提出,使得关于属性约简的研究不断趋于完善.在这些众多的属性约简方法中,数据之间关系的描述仍然建立在传统的明确集基础上,在属性约简的过程中,目标对象通过明确集来进行近似逼近,这样便不能对属性达到更好的属性重要度度量效果^[20].

直觉模糊集^[21]是在传统模糊集基础上的进一步推广.近年来,基于直觉模糊集的粗糙集模型^[22]被提出,在该模型中,目标的近似通过论域的直觉模糊关系粒化来描述,使得达到了更优越的粗糙逼近效果,因而基于直觉模糊粗糙集直觉模糊粒化的研究也逐渐趋热^[23-24],针对直觉模糊粗糙集的直觉模糊粒化,目前学者们提出了大量的属性约简算法.例如路艳平等学者^[25]运用直觉模糊粗糙集的直觉模糊粒化去构造正区域,并基于正区域进行属性约简.陈昊等学者^[26]运用属性间互信息的依赖度进行属性约简.王学明等学者^[27]根据粒化结果后的隶属度和非隶属度分别定义了依赖度和非依赖度的概念,并用于属性约简算法的构建.鲍忠奎等学者^[28]运用直觉模糊粒化的正区域提出优势关系下的属性约简方法.可以看出,目前基于直觉模糊粒化的属性约简方法大多是基于传统的正区域或依赖度的方法构建.

信息系统中的信息熵理论是度量属性重要度的一种重要方法^[8-9,29].为了对信息系统的属性达到更为优越的重要度度量,本文在直觉模糊关系的粒化结果基础上,根据隶属度和非隶属度定义了两种信息熵的表示方式,然后将这两种信息熵合并作为整个信息系统的信息熵度量,最后基于该度量构造了相应的属性约简算法,实验结果表明,该算法选择出的约简集更具优越,并且算法有着较高的约简效率,

在实际中运用于数据集的属性约简是有效的.

2 基本理论

定义 1 设信息系统 $IS = (U, AT)^{[1]}$, 其中 U 称为论域, AT 为属性集, 对于 $A \subseteq AT$ 在 U 上诱导的等价关系 R_A 定义为

$$R_A = \{(x, y) \in U \times U \mid \forall a \in A, a(x) = a(y)\} \quad (1)$$

同时, 对象 $x \in U$ 在 R_A 下的等价类定义为 $[x]_{R_A}$, 由 R_A 诱导的划分为 U/R_A , 根据粒计算的视角, 这里的划分可以看成是等价关系对信息系统论域生成的一组等价关系粒化.

定义 2 对于 $IS = (U, AT)^{[2]}$, $A \subseteq AT$, 定义 $X \subseteq U$ 基于等价关系 R_A 的下近似和上近似分别为

$$R_A(X) = \{x \in U \mid [x]_{R_A} \subseteq X\} \quad (2)$$

$$\bar{R}_A(X) = \{x \in U \mid [x]_{R_A} \cap X \neq \emptyset\} \quad (3)$$

称 $\langle R_A(X), \bar{R}_A(X) \rangle$ 为 X 在 R_A 下的粗糙集.

定义 3 考虑论域 $U = \{x_1, x_2, \dots, x_n\}^{[21]}$, U 上的直觉模糊集 $\hat{A}$ 表示为

$$\hat{A} = \{\langle x_i, \mu_{\hat{A}}(x_i), \gamma_{\hat{A}}(x_i) \rangle \mid 1 \leq i \leq n\} \quad (4)$$

式中, $\mu_{\hat{A}}(x_i): U \rightarrow [0, 1]$, $\gamma_{\hat{A}}(x_i): U \rightarrow [0, 1]$, 并且 $\forall x_i \in U$ 满足 $\mu_{\hat{A}}(x_i) + \gamma_{\hat{A}}(x_i) \in [0, 1]$. 其中 $\mu_{\hat{A}}(x_i)$ 和 $\gamma_{\hat{A}}(x_i)$ 分别称为 x_i 关于直觉模糊集 $\hat{A}$ 的隶属度和非隶属度, 同时 $\pi_{\hat{A}}(x_i) = 1 - \mu_{\hat{A}}(x_i) - \gamma_{\hat{A}}(x_i)$ 称为对象 x_i 关于直觉模糊集 $\hat{A}$ 的犹豫度. 论域 U 上直觉模糊集的全体表示为 $IF(U)$. 显然, 模糊集是直觉模糊集的特例.

定义 4 设论域 U 上的两个直觉模糊集 $\hat{A}$ 和 $\hat{B}$ 分别^[21]为

$$\hat{A} = \{\langle x_i, \mu_{\hat{A}}(x_i), \gamma_{\hat{A}}(x_i) \rangle \mid 1 \leq i \leq n\} \quad (5)$$

$$\hat{B} = \{\langle x_i, \mu_{\hat{B}}(x_i), \gamma_{\hat{B}}(x_i) \rangle \mid 1 \leq i \leq n\} \quad (6)$$

那么

(1) $\hat{A}$ 的补集为

$$\sim \hat{A} = \{\langle x_i, \gamma_{\hat{A}}(x_i), \mu_{\hat{A}}(x_i) \rangle \mid 1 \leq i \leq n\};$$

$$(2) \hat{A} = \hat{B} \Leftrightarrow \mu_{\hat{A}}(x) = \mu_{\hat{B}}(x) \wedge \gamma_{\hat{A}}(x) = \gamma_{\hat{B}}(x);$$

$$(3) \hat{A} \subseteq \hat{B} \Leftrightarrow \mu_{\hat{A}}(x) \leq \mu_{\hat{B}}(x) \wedge \gamma_{\hat{A}}(x) \geq \gamma_{\hat{B}}(x);$$

$$(4) \hat{A} \cup \hat{B} = \{ \langle x_i, \max(\mu_{\hat{A}}(x), \mu_{\hat{B}}(x)), \min(\gamma_{\hat{A}}(x), \gamma_{\hat{B}}(x)) \rangle \mid 1 \leq i \leq n \};$$

$$(5) \hat{A} \cap \hat{B} = \{ \langle x_i, \min(\mu_{\hat{A}}(x), \mu_{\hat{B}}(x)), \max(\gamma_{\hat{A}}(x), \gamma_{\hat{B}}(x)) \rangle \mid 1 \leq i \leq n \}.$$

定义 5 对于信息系统^[22] $IS = (U, AT)$, $A \subseteq AT$, 定义 U 上的直觉模糊关系 $\tilde{R}_A$ 为

$$\tilde{R}_A = \{ \langle (x, y), \mu_{\tilde{R}_A}(x, y), \gamma_{\tilde{R}_A}(x, y) \rangle \mid (x, y) \in U \times U \} \quad (7)$$

式中, $\mu_{\tilde{R}_A}, \gamma_{\tilde{R}_A} \in [0, 1]$; $\mu_{\tilde{R}_A} + \gamma_{\tilde{R}_A} \in [0, 1]$. 并且满足 $\mu_{\tilde{R}_A}(x, x) = 1, \gamma_{\tilde{R}_A}(x, x) = 0$. 对于 $\forall a \in A, \tilde{R}_A = \bigcap \tilde{R}_{\{a\}}$. 并且直觉模糊关系矩阵定义为

$$M_{\tilde{R}_A}^{n \times n} = (\langle \mu_{\tilde{R}_A}(x_i, y_j), \gamma_{\tilde{R}_A}(x_i, y_j) \rangle)_{n \times n} \quad (8)$$

定义 6 对于信息系统 $IS = (U, AT)$, $A \subseteq AT$ 诱导的直觉模糊关系为 $\tilde{R}_A, \hat{X} \subseteq IF(U)$ 基于 $\tilde{R}_A$ 的下近似和上近似分别定义为

$$\tilde{R}_A(\hat{X}) = \{ \langle x, \mu_{\tilde{R}_A(\hat{X})}(x), \gamma_{\tilde{R}_A(\hat{X})}(x) \rangle \mid x \in U \} \quad (9)$$

$$\bar{\tilde{R}}_A(\hat{X}) = \{ \langle x, \bar{\mu}_{\tilde{R}_A(\hat{X})}(x), \bar{\gamma}_{\tilde{R}_A(\hat{X})}(x) \rangle \mid x \in U \} \quad (10)$$

式中

$$\begin{aligned} \mu_{\tilde{R}_A(\hat{X})}(x) &= \bigwedge_{y \in U} (\gamma_{\tilde{R}_A}(x, y) \vee \mu_{\hat{X}}(y)); \\ \gamma_{\tilde{R}_A(\hat{X})}(x) &= \bigvee_{y \in U} (\mu_{\tilde{R}_A}(x, y) \wedge \gamma_{\hat{X}}(y)); \\ \bar{\mu}_{\tilde{R}_A(\hat{X})}(x) &= \bigvee_{y \in U} (\mu_{\tilde{R}_A}(x, y) \wedge \mu_{\hat{X}}(y)); \\ \bar{\gamma}_{\tilde{R}_A(\hat{X})}(x) &= \bigwedge_{y \in U} (\gamma_{\tilde{R}_A}(x, y) \vee \gamma_{\hat{X}}(y)). \end{aligned}$$

称 $(\tilde{R}_A(\hat{X}), \bar{\tilde{R}}_A(\hat{X}))$ 为 $\hat{X}$ 在 $\tilde{R}_A$ 下的直觉模糊粗糙集.

3 直觉模糊关系粒化的信息熵属性约简算法

3.1 直觉模糊关系粒化的信息熵

在粒计算理论中, 梁吉业等学者^[29]在信息系统中提出一种信息熵的不确定性度量方法, Qian 等^[30]学者在此基础上将信息熵进一步推广, 提出了基于模糊粗糙集的信息熵度量.

定义 7 设信息系统 $IS = (U, AT)$, 令 $|U| = n, A \subseteq AT$ 诱导出的模糊相似关系为 R_A , 并且论域

在 R_A 上的模糊信息粒化为

$$K(R_A) = \{S_A(x_1), S_A(x_2), \dots, S_A(x_n)\} \quad (11)$$

式中, $S_A(x_i)$ 称为 $x_i \in U$ 在模糊相似关系 R_A 下的模糊信息粒, $S_A(x_i)$ 是一个模糊集, 具体可表示为

$$S_A(x_i) = \frac{R_A(x_i, x_1)}{x_1} + \frac{R_A(x_i, x_2)}{x_2} + \dots + \frac{R_A(x_i, x_n)}{x_n}$$

定义信息系统在模糊相似关系 R_A 上的模糊信息熵^[30] 定义为

$$FE(R_A) = \frac{1}{n} \sum_{i=1}^n (1 - \frac{|S_A(x_i)|}{n}) \quad (12)$$

式中, $|\cdot|$ 表示模糊集的基数.

在定义 7 的基础上, 可以将信息熵在直觉模糊集上进行进一步推广.

由于在直觉模糊集中, 对象与集合之间的关系通过多个参数来刻画, 因此 Szmidt 等学者^[31] 和 Vlachos 等学者^[32] 在直觉模糊集中分别定义了直觉模糊集集合基数的表示方法.

定义 8 对于论域 U 上的直觉模糊集 $\hat{A} = \{ \langle x_i, \mu_{\hat{A}}(x_i), \gamma_{\hat{A}}(x_i) \rangle \mid 1 \leq i \leq n \}$, 由隶属度 $\mu_{\hat{A}}(x)$ 确定的直觉模糊集基数定义为 $|\hat{A}|^+ = \sum_{i=1}^n \mu_{\hat{A}}(x_i)$, 由非隶属度 $\gamma_{\hat{A}}(x)$ 确定的直觉模糊集基数定义为 $|\hat{A}|^- = \sum_{i=1}^n \gamma_{\hat{A}}(x_i)$.

例 1. 设论域 $U = \{x_1, x_2, x_3, x_4\}$ 上的直觉模糊集 $\hat{A}$ 为 $\hat{A} = \{ \langle 0.6, 0.2 \rangle, \langle 0.4, 0.3 \rangle, \langle 0.7, 0.1 \rangle, \langle 0.2, 0.3 \rangle \}$.

那么根据定义 8 有

$$|\hat{A}|^+ = 0.6 + 0.4 + 0.7 + 0.2 = 1.9,$$

$$|\hat{A}|^- = 0.2 + 0.3 + 0.1 + 0.3 = 0.9.$$

根据这两种直觉模糊集的基数表示方式, 这里可以得到基于隶属关系和非隶属关系的直觉模糊信息熵.

Qian 等^[30]学者基于模糊相似关系, 提出了模糊关系的信息熵模型, 模糊相似关系描述了对对象与集合之间的隶属程度, 而在直觉模糊相似关系中, 对象与集合之间的隶属程度通过隶属度和非隶属度两个方面来描述, 因此可以将 Qian 等学者的信息熵模型在直觉模糊关系下进行推广, 提出直觉模糊信息熵模型.

定义 9 对于信息系统 $IS = (U, AT)$, 令 $|U| = n, A \subseteq AT$ 诱导的直觉模糊关系为 $\tilde{R}_A$, 那么论域在 $\tilde{R}_A$ 下的直觉模糊粒化定义为

$$K(\tilde{R}_A) = \{S_{\tilde{R}_A}(x_1), S_{\tilde{R}_A}(x_2), \dots, S_{\tilde{R}_A}(x_n)\} \quad (13)$$

式中, $S_{\tilde{R}_A}(x_i)$ 表示 $x_i \in U$ 在 $\tilde{R}_A$ 的直觉模糊信息粒, 表示为

$$S_{\tilde{R}_A}(x_i) = \{\langle x_j, \mu_{\tilde{R}_A}(x_i, x_j), \gamma_{\tilde{R}_A}(x_i, x_j) \rangle \mid 1 \leq j \leq n\} \quad (14)$$

类似于 Qian 等学者的模糊粒化, 接下来将通过信息系统的直觉模糊粒化去构造信息熵。

定义 10 对于信息系统 $IS = (U, AT)$, 令 $|U| = n, A \subseteq AT$ 诱导的直觉模糊关系为 $\tilde{R}_A$, 那么基于 $\tilde{R}_A$ 的隶属和非隶属直觉模糊信息熵分别定义为

$$IFE^+(\tilde{R}_A) = \frac{1}{n} \sum_{i=1}^n (1 - \frac{|S_{\tilde{R}_A}(x_i)|^+}{n}) \quad (15)$$

$$IFE^-(\tilde{R}_A) = \frac{1}{n} \sum_{i=1}^n (1 - \frac{|S_{\tilde{R}_A}(x_i)|^-}{n}) \quad (16)$$

在定义 10 中, 这两种信息熵是通过不同的直觉模糊隶属角度描述信息系统的不确定性, 其中 $IFE^+(\tilde{R}_A)$ 表达了直觉模糊隶属度视角的熵度量, $IFE^-(\tilde{R}_A)$ 表达了直觉模糊非隶属度视角的熵度量, 因而考虑将它们结合作为整个直觉模糊集的信息熵度量。

定义 11 对于信息系统 $IS = (U, AT)$, 令 $|U| = n, A \subseteq AT$ 诱导的直觉模糊关系为 $\tilde{R}_A$, 基于 $\tilde{R}_A$ 的直觉模糊信息熵定义为

$$IFE(\tilde{R}_A) = IFE^+(\tilde{R}_A) + IFE^-(\tilde{R}_A) \quad (17)$$

性质 1 对于信息系统 $IS = (U, AT)$, 令 $|U| = n, A \subseteq AT$ 诱导的直觉模糊关系为 $\tilde{R}_A$, 直觉模糊信息熵满足

$$1 \leq IFE(\tilde{R}_A) \leq 2 \quad (18)$$

证明: 将定义 10 中的隶属和非隶属直觉模糊信息熵代入定义 11 可以得到

$$\begin{aligned} IFE(\tilde{R}_A) &= IFE^+(\tilde{R}_A) + IFE^-(\tilde{R}_A) \\ &= \sum_{i=1}^n \frac{2n - (|S_{\tilde{R}_A}(x_i)|^+ + |S_{\tilde{R}_A}(x_i)|^-)}{n^2} \end{aligned}$$

令 $|S_{\tilde{R}_A}(x_i)|^+ = x, |S_{\tilde{R}_A}(x_i)|^- = y$, 定义函数

$$f(x, y) = 2n - (x + y),$$

$$\text{式中, 约束条件为} \begin{cases} 0 \leq x + y \leq 1 \\ x \geq 0 \\ y \geq 0 \end{cases}.$$

分析 $f(x, y)$ 在约束条件下的极值. 采用线性规划的方法进行求解, 即求 $y = -x + 2n - f(x, y)$ 在图 1 中阴影区域的截距范围。

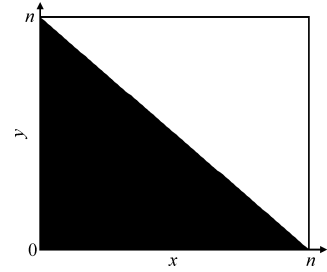


图 1 线性规划约束范围

因此有 $0 \leq 2n - f(x, y) \leq n$, 所以 $n \leq f(x, y) \leq 2n$.

因而 $\sum_{i=1}^n \frac{n}{n^2} \leq IFE(\tilde{R}_A) \leq \sum_{i=1}^n \frac{2n}{n^2}$, 因此满

足关系 $1 \leq IFE(\tilde{R}_A) \leq 2$. 证毕。

性质 2 对于信息系统 $IS = (U, AT)$, 令 $|U| = n, B \subseteq A \subseteq AT$, 并且 B, A 诱导的直觉模糊关系分别为 $\tilde{R}_B, \tilde{R}_A$. 那么直觉模糊信息熵满足关系

$$IFE(\tilde{R}_A) \leq IFE(\tilde{R}_B) \quad (19)$$

证明: 由性质 1 的证明可知, 那么随着 $|S_A(x_i)|^+ + |S_A(x_i)|^-$ 增大, $IFE(\tilde{R}_A)$ 是逐渐减小的. 因此当 $B \subseteq A$, 根据定义 5 有 $\tilde{R}_A \subseteq \tilde{R}_B$, 因此 $|S_A(x_i)|^+ + |S_A(x_i)|^- \leq |S_B(x_i)|^+ + |S_B(x_i)|^-$, 所以性质 2 成立, 证毕。

性质 2 表明, 随着属性集的增加, 信息系统的直觉模糊信息熵值单调不减, 根据这一性质可以导出信息系统的属性约简算法。

3.2 属性约简算法

在本节将提出直觉模糊信息熵的属性约简算法。

定义 12 对于信息系统 $IS = (U, AT)$, $A \subseteq AT$ 诱导的直觉模糊关系为 $\tilde{R}_A$, 若直觉模糊信息熵满足 $IFE(\tilde{R}_A) = IFE(\tilde{R}_{AT})$, 并且对于 $\forall a \in A$, 都有 $IFE(\tilde{R}_{A-\{a\}}) > IFE(\tilde{R}_{AT})$. 那么称属性集 A 为信息系统的约简集。

通过定义 12 可以看出, 如果属性子集 $A \subseteq AT$ 为信息系统的约简集, 首先要满足 $\tilde{R}_A$ 的直觉模糊信

息熵和全属性集的直觉模糊信息熵相等,然后保证该属性子集是所有满足条件中最小的. 基于该属性约简方法,这里给出基于直觉模糊信息熵的属性重要度度量方法.

定义 13 对于信息系统 $IS = (U, AT), A \subseteq AT$, 诱导的直觉模糊关系分别为 $\tilde{R}_A, \tilde{R}_{AT}$. 这里定义两种属性重要度的表示方式, 分别表示为内属性重要度 sig^{in} 和外属性重要度 sig^{out} .

对于 $\forall a \in A$ 相对于 A 的直觉模糊信息熵内属性重要度 $\text{sig}_A^{\text{in}}(a)$ 定义为

$$\text{sig}_A^{\text{in}}(a) = \text{IFE}(\tilde{R}_{A-\{a\}}) - \text{IFE}(\tilde{R}_A). \tag{20}$$

对于 $\forall a \in AT - A$ 相对于 A 的直觉模糊信息熵外属性重要度 $\text{sig}_A^{\text{out}}(a)$ 定义为

$$\text{sig}_A^{\text{out}}(a) = \text{IFE}(\tilde{R}_A) - \text{IFE}(\tilde{R}_{A \cup \{a\}}). \tag{21}$$

基于定义 12 和定义 13, 这里提出一种基于直觉模糊信息熵的信息系统属性约简算法.

算法 1 计算直觉模糊信息熵.

输入: 信息系统 $IS = (U, AT), A \subseteq AT$,

输出: $\text{IFE}(\tilde{R}_A)$.

Step1. 初始化 $\text{IFE}^+(\tilde{R}_A) = \text{IFE}^-(\tilde{R}_A) = 0$.

Step2. 计算 U 中所有对象的直觉模糊关系矩阵为

$$M_{R_A}^{n \times n} = (\langle \mu_{\tilde{R}_A}(x_i, y_j), \gamma_{\tilde{R}_A}(x_i, y_j) \rangle)_{n \times n}$$

Step3. 计算每个对象 x_i 的 $|S_A(x_i)|^+$ 和 $|S_A(x_i)|^-$. / * 根据定义 8 * /

Step4. 返回 $\text{IFE}(\tilde{R}_A)$, 算法终止. / * 根据定义 10 和 11 * /

算法 1 的计算量主要集中在 Step2 和 Step3, 设 $|A| = a$, 因此时间复杂度为 $O(a \cdot n^2)$. 在算法 1 的基础上, 利用启发式搜索的方式构造直觉模糊信息熵的属性约简算法.

算法 2 属性约简算法.

输入: 信息系统 $IS = (U, AT)$.

输出: 约简集 reduct .

Step1. 初始化约简集 $\text{reduct} \leftarrow \emptyset, \text{IFE}(\tilde{R}_{\emptyset}) = 2$.

Step2. 对于 $\forall a \in AT$, 根据计算定义 13 计算出 a 的属性重要度 $\text{sig}_{AT}^{\text{in}}(a)$. 若 $\text{sig}_{AT}^{\text{in}}(a) > 0$, 那么 $\text{reduct} \leftarrow \text{reduct} \cup \{a\}$.

Step3. 如果 $\text{IFE}(\tilde{R}_{\text{reduct}}) = \text{IFE}(\tilde{R}_{AT})$, 那么进入 Step5, 否则进入 Step4.

Step4. 对于 $\forall a \in AT - \text{reduct}$, 计算属性重要

度 $\text{sig}_{\text{reduct}}^{\text{out}}(a)$, 并选择出所有属性重要度最大的属性, 加入到 reduct 中. 重复进行 Step4, 直到满足 $\text{IFE}(\tilde{R}_{\text{reduct}}) = \text{IFE}(\tilde{R}_{AT})$.

Step5. 对于 $\forall a \in \text{reduct}$, 找出使等式 $\text{IFE}(\tilde{R}_{\text{reduct}-\{a\}}) = \text{IFE}(\tilde{R}_{AT})$ 成立的属性, 并将这些属性的集合记为 Ω . 若 $\Omega = \emptyset$, 进入 Step7, 否则进入 Step6.

Step6. 对于 Step5 中的属性集 Ω , 任意选择一个属性 b , 进行 $\text{reduct} \leftarrow \text{reduct} - \{b\}$, 判断 $\text{IFE}(\tilde{R}_{\text{reduct}}) = \text{IFE}(\tilde{R}_{AT})$ 是否成立, 若成立则进入 Step5, 否则进入 Step7.

Step7. 返回约简集 reduct , 算法终止.

在算法 2 中, 设 $|AT| = c, |U| = n$, 根据文献 [33-34] 的时间复杂度分析, 算法 2 的时间复杂度为 $O(c^2 \cdot n^2)$.

4 实验分析

通过实验来验证本文所提出属性约简算法的优越性. 实验中所采用的 UCI 数据集如表 1 所示, 参与对比的算法分别为 (1) 基于邻域熵的属性约简^[5], 记为 AR-1 算法, (2) 相对决策熵的属性约简算法^[24], 记为 AR-2 算法, (3) 基于模糊关系粒化的信息熵属性约简算法^[11], 记为 AR-3 算法, (4) 基于直觉模糊粗糙集的依赖度属性约简算法^[26], 记为 AR-4 算法, (5) 基于直觉模糊粗糙集的依赖度与非依赖度双重度量属性约简算法^[27], 记为 AR-5 算法, 另外记本文的直觉模糊粒化信息熵的属性约简为 AR-6 算法. 本文算法中直觉模糊关系的构造采用文献[35]提出的方法, 即通过证据理论中信任函数和似然函数等函数去度量直觉模糊关系中的隶属度和非隶属度.

表 1 UCI 数据集

编号	数据集	对象	属性	类
1	glass	214	9	6
2	wine	178	13	3
3	segment	2 310	19	7
4	wdbc	569	31	2
5	biodeg	1 055	41	2
6	sonar	208	60	2
7	gearbox	1 603	72	4

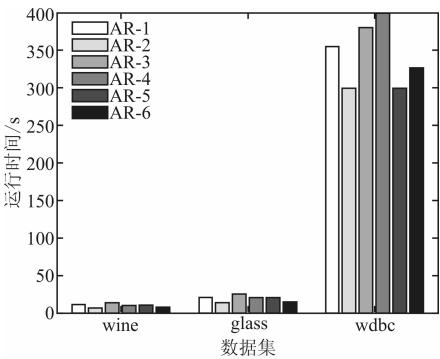
实验所采用的硬件环境为 Intel(R) Core(TM) i5 6500, 3.2 GHz CPU 和 8GB 内存的 Windows10 操作系统个人主机, 所有算法采用 VC++6.0 进行编程.

实验总共分为两部分,在第一部分中,让 6 种算法分别对表 1 中数据集进行属性约简,然后对得到的约简结果采用支持向量机(SVM)分类器和 k 近邻(kNN)分类器分别进行分类训练,从而得到约简集对应的分类精度.其中 kNN 分类器需要确定 k 值,本实验采用交叉验证的方法来确定各个数据集的最优 k 值,计算得到的结果分别为:glass 最优 k 值为 3,wine 最优 k 值为 7,segment 最优 k 值为 15,wdbc 最优 k 值为 5,biodeg 最优 k 值为 7,sonar 最优 k 值为 3,gearbox 最优 k 值为 11.所有数据集在 6 种算法下约简集大小和分类精度分别如表 2、表 3 和表 4 所示,其中“√”表示表格中每行数值的最小值.实验的第二部分是关于 6 种算法属性约简效率的比较,分别让 6 种算法对各个数据集重复属性约简 10 次,取 10 次约简时间消耗的平均值作为算法在该数据集上的运行时间,所有实验结果如图 2 所示.

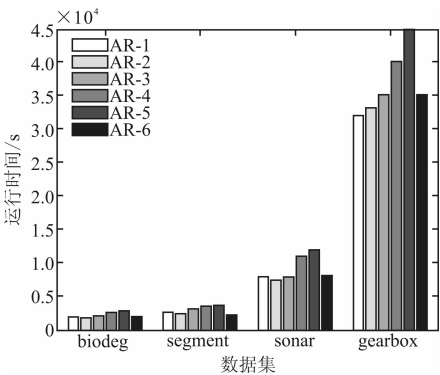
表 2 约简集大小比较							
数据集	原始属性集	AR-1	AR-2	AR-3	AR-4	AR-5	AR-6
glass	9	7	6	6	6	5√	5√
wine	13	7	6	7	5√	5√	5√
segment	19	9	10	9	7√	8	7√
wdbc	31	12	12	11	10	10	9√
biodeg	41	15	14	14	13	12	11√
sonar	60	17	17	15	13	12√	13
gearbox	72	16	15	15	13	13	12√

表 3 SVM 分类精度比较								%
数据集	原始属性集	AR-1	AR-2	AR-3	AR-4	AR-5	AR-6	
glass	58.425 9	59.345 6	60.534 7	60.534 7	60.534 7	61.354 2√	61.354 2√	
wine	97.684 2	98.542 7	98.127 9	98.542 7	99.142 0√	99.142 0√	99.142 0√	
segment	92.275 1	96.242 3	96.242 3	95.839 7	95.488 4	95.037 5	96.756 2√	
wdbc	96.648 7	95.724 5	96.248 9	98.435 7√	96.547 5	95.842 7	97.385 2	
biodeg	85.634 8	86.478 4	86.478 4	85.423 1	86.564 2	86.765 2	87.232 8√	
sonar	87.389 2	85.685 2	86.934 5√	87.348 5	86.637 5	86.214 3	86.637 5	
gearbox	95.462 3	97.643 8	95.972 5	96.462 8	97.824 7	97.824 3	98.173 5√	

表 4 kNN 分类精度比较								%
数据集	原始属性集	AR-1	AR-2	AR-3	AR-4	AR-5	AR-6	
glass	60.895 2	61.248 2	61.756 4	61.756 4	61.756 4	62.342 8√	62.342 8√	
wine	96.456 7	97.349 4	96.930 4	97.349 4	97.830 5√	97.830 5√	97.830 5√	
segment	94.842 4	95.623 4	95.623 4	97.678 3√	96.384 2	96.633 4	97.512 8	
wdbc	97.452 8	97.643 1	96.682 4	96.463 4	98.460 1	97.124 9	98.730 4√	
biodeg	83.947 2	84.015 7	84.015 7	85.722 4	86.743 2√	85.367 0	85.675 2	
sonar	88.534 6	90.456 2	89.484 2	89.315 9	91.215 8√	90.562 4	91.215 8√	
gearbox	97.455 7	99.128 5	99.262 1	97.382 4	98.027 8	98.638 2	99.834 4√	



(a) 部分数据集运行时间比较



(b) 部分数据集运行时间比较

图 2 运行时间比较

在表 2 中,“√”表示 6 种算法的最小约简集.观察表 2 可以发现,AR-4,AR-5 和 AR-6 算法属性约简得到的约简集大小均小于 AR-1,AR-2 和 AR-3 算法的约简集,这是由于 AR-4,AR-5 和 AR-6 算法均以直觉模糊关系粒化为基础构造的属性约简,而直觉模糊粒化通过隶属度和非隶属度两种参数来描述数据之间的相似性,对属性有着更好的评估度量效果,从而达到更好的属性选择能力.对于 AR-4,AR-5 和 AR-6 算法,本文提出的 AR-6 算法约简出的属性子集更加的小,这主要是由于 AR-4 和 AR-5 算法采用的是属性依赖度的方法构造属性约简,而 AR-6 算法采用隶属度和非隶属度分别对目标对象进行信息熵度量,在属性的评估方面有着更高的性能,因而约简出更有效的属性子集.

表 3 和表 4 分别表示的是 SVM 分类器和 kNN 分类器对表 2 中属性子集的分类精度结果,其中“√”表示 6 种算法约简结果分类精度的最大值.观察可以发现,基于直觉模糊粒化构造的 AR-4,AR-5 和 AR-6 算法,约简出的属性子集整体上具有更高的分类精度,原因同样是由于直觉模糊关系在数据相似性描述的优越性.同时在大部分数据集中,AR-6 算法约简出的属性子集具有更高的分类精度,这

说明了在直觉模糊关系中,信息熵的度量比依赖度的度量更加的有效。

在图 2 所示的算法运行时间比较结果中,当数据集的规模较小时,例如 wine, glass, wdbc, bio-deg 和 segment 数据集,这 6 种算法的属性约简耗时相近,但是当数据集的规模增大时,例如 sonar 和 gearbox 数据集,AR-4 和 AR-5 算法的约简耗时出现明显的增加,其余算法的约简耗时保持相近,这一方面说明了基于直觉模糊粒化的依赖度属性约简不适宜规模较大的数据集,另一方面表明本文所提出的直觉模糊关系粒化信息熵的属性约简算法有着较高的算法效率。

综合以上的实验结果表明,本文所提出的直觉模糊粒化信息熵的属性约简算法在约简结果和算法效率方面都具有更好的优越性。

5 结束语

本文在直觉模糊关系的基础上对信息系统进行直觉模糊粒化,并根据粒化结果分别采用隶属度和非隶属度的概念定义了两种信息熵形式,然后将这两种信息熵结合起来作为整体信息系统的信息熵度量,并基于该度量作为信息系统中属性重要度的评估,最后提出了一种属性约简算法,仿真实验表明了该算法的优越性。在将来的研究中,将进一步探索混合类型数据集的约简问题。

参考文献:

- [1] HU J, WANG G Y, ZHANG Q H, et al. Attribute reduction based on granular computing[J]. Lecture Notes in Computer Science, 2006, 4259(3):458-466.
- [2] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.
- [3] ZADEH L. A fuzzy set theoretic interpretation of linguistic hedges[J]. Journal of Cybernetics, 1972, 2(3):4-34.
- [4] 向伟. 邻域系统中对象变化的动态属性约简算法[J]. 计算机应用与软件, 2018, 35(11): 278-282.
- [5] 王光琼. 基于邻域组合熵的属性约简算法[J]. 计算机应用与软件, 2018, 35(12): 269-273.
- [6] 吕洁, 刘利民, 胡皎月, 等. 基于 MapReduce 的高效粗糙集属性约简算法[J]. 计算机应用与软件, 2017, 34(4): 272-276.
- [7] QIAN Y H, LIANG J Y, PEDRYCZ W, et al. Positive approximation: An accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9/10):597-618.
- [8] DAI J H, WANG W T, TIAN H W, et al. Attribute selection based on a new conditional entropy for incomplete decision systems[J]. Knowledge-Based Systems, 2013, 39(2): 207-213.
- [9] 续欣莹, 刘海涛, 谢珺, 等. 信息观下基于不一致邻域矩阵的属性约简[J]. 控制与决策, 2016, 31(1): 130-136.
- [10] LIN Y J, LI J J, LIN P R, et al. Feature selection via neighborhood multi-granulation fusion[J]. Knowledge-Based Systems, 2014, 67(3):162-168.
- [11] ZHANG X, MEI C L, CHEN D G, et al. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition, 2016, 56(1):1-15.
- [12] WANG C Z, QI Y L, SHAO M W, et al. A fitting model for feature selection with fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(4): 741-753.
- [13] MA X A, WANG G Y, YU H, et al. Decision region distribution preservation reduction in decision-theoretic rough set model[J]. Information Sciences, 2014, 278(6): 614-640.
- [14] JING Y G, LI T R, HANG J F, et al. An incremental attribute attribute reduction approach based on knowledge granularity under the attribute generalization[J]. International Journal of Approximate Reasoning, 2016, 76(2):80-95.
- [15] JING Y G, LI T R, HANG J F, et al. A group incremental reduction algorithm with varying data values[J]. International Journal of Intelligent Systems, 2017, 32(9):900-925.
- [16] JING Y G, LI T R, HAMIDO F, et al. An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view[J]. Information Sciences, 2017, 411(5):23-38.
- [17] JING Y G, LI T R, LUO C, et al. An incremental approach for attribute reduction based on knowledge granularity[J]. Knowledge-based Systems, 2016(104):24-38.
- [18] CHEN H M, LI T R, CAI Y, et al. Parallel attribute reduction in dominance-based neighborhood rough set[J]. Information Sciences, 2016, 373(10):351-368.
- [19] ZHANG J B, LI T R, PAN Y. PLAR: Parallel large-scale attribute reduction on cloud systems[C]// International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE, 2014: 184-191.

- [20] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2):161-166.
- [21] ATANASSOV K. Intuitionistic fuzzy sets[J]. Fuzzy Sets & Systems, 1986, 20(1):87-96.
- [22] ZHOU L, WU W Z. Characterization of rough set approximations in Atanassov intuitionistic fuzzy set theory[J]. Computers & Mathematics with Applications, 2011, 62(1):282-296.
- [23] ZHANG Z M. A rough set approach to intuitionistic fuzzy soft set based decision making [J]. Applied Mathematical Modelling, 2012, 36(10):4605-4633.
- [24] 薛占熬, 司小朦, 袁艺林, 等. 多粒度邻域粗糙直觉模糊集模型[J]. 模式识别与人工智能, 2017, 30(1):11-20.
- [25] 路艳丽, 雷英杰, 华继学. 基于直觉模糊粗糙集的属性约简[J]. 控制与决策, 2009, 24(3):335-341.
- [26] 陈昊, 杨会成. 一种直觉模糊粗糙集属性约简新算法[J]. 小型微型计算机系统, 2011, 32(3):506-510.
- [27] 王学明, 舒兰. 基于直觉模糊相似关系下的粗糙集属性约简[J]. 模糊系统与数学, 2012, 26(5):185-189.
- [28] 鲍忠奎, 杨善林. 直觉模糊目标信息系统的正域约简[J]. 中国科学技术大学学报, 2015(4): 329-336.
- [29] 梁吉业, 钱宇华. 信息系统中的信息粒与熵理论[J]. 中国科学 E 辑:信息科学, 2008, 38(12): 2048-2065.
- [30] QIAN Y H, LIANG J Y, WU W Z, et al. Information granularity in fuzzy binary GrC model[J]. IEEE Transactions on Fuzzy Systems, 2011, 19 (2): 253-264.
- [31] SZMIDT E, KACPRZYK J. Entropy for intuitionistic fuzzy sets[J]. Fuzzy Sets & Systems, 2001, 118(3): 467-477.
- [32] VLACHOS I, SERGIADIS G. Subsethood, entropy, and cardinality for interval-valued fuzzy sets-An algebraic derivation[J]. Fuzzy Sets & Systems, 2007, 158 (12):1384-1396.
- [33] 覃丽珍, 姚炳学, 李金海. 基于信息量的完备覆盖约简算法[J]. 计算机科学, 2012, 39(10):235-239.
- [34] JIANG F, SUI Y, ZHOU L. A relative decision entropy-based feature selection approach [J]. Pattern Recognition, 2015, 48(7):2151-2163.
- [35] 邢清华, 刘付显. 直觉模糊集隶属度与非隶属度函数的确定方法[J]. 控制与决策, 2009, 24(3):393-397

作者简介:

陈曦女, (1982-), 博士研究生, 讲师. 研究方向为智能计算、项目管理和风险管理.

E-mail: chenxi198211@126.com.