

引用格式:刘正,陈雪勤,张书锋. 基于最小化邻域互信息的邻域熵属性约简算法[J]. 微电子学与计算机, 2020, 37(3): 26-32.
[LIU Z, CHEN X Q, ZHANG S F. Neighborhood entropy attribute reduction based on minimizing neighborhood mutual information [J]. Microelectronics & Computer, 2020, 37(3): 26-32.]

基于最小化邻域互信息的邻域熵属性约简算法

刘 正^{1,2}, 陈雪勤¹, 张书锋²

(1 苏州大学 电子信息学院, 江苏 苏州 215006;

2 苏州工业园区服务外包职业学院 信息工程学院, 江苏 苏州 215000)

摘 要: 属性约简是粗糙集理论的重要研究内容, 混合型信息系统下的属性约简是目前的主要研究方向. 在邻域粗糙集模型中, 基于邻域熵的方法在进行属性约简时, 由于未考虑属性之间的独立性, 本文通过融入邻域互信息熵提出一种改进的属性约简算法. 文中首先在邻域熵的基础上, 提出了混合型信息系统下的邻域互信息熵, 然后通过理论分析表明了邻域互信息熵可以作为属性之间独立性的评估, 最后将邻域互信息熵融入传统的邻域熵属性约简中, 提出一种基于最小化邻域互信息的邻域熵属性约简算法. 仿真实验结果表明, 该算法可以进一步地提高属性约简结果中属性的独立程度, 比相关的属性约简算法具有更高的约简性能.

关键词: 粗糙集; 属性约简; 混合型信息系统; 邻域熵; 邻域互信息熵

中图分类号: TP18

文献标识码: A

文章编号: 1000—7180(2020)03—0026—07

Neighborhood entropy attribute reduction based on minimizing neighborhood mutual information

LIU Zheng^{1,2}, CHEN Xue-qin¹, ZHANG Shu-feng²

(1 College of Electronic Information, Suzhou University, Suzhou 215006, China;

2 College of Information Engineering, Service Outsourcing Vocational College of
Suzhou Industrial Park, Suzhou 215000, China)

Abstract: Attribute reduction is an important research content of rough set theory. Attribute reduction in hybrid information system is the main research direction at present. In the neighborhood rough set model, an improved attribute reduction algorithm based on neighborhood entropy is proposed by incorporating neighborhood mutual information entropy, because the independence between attributes is not considered. Firstly, based on the neighborhood entropy, the neighborhood mutual information entropy in hybrid information system is proposed. Then the theoretical analysis shows that the neighborhood mutual information entropy can be used to evaluate the independence of attributes. Finally, the neighborhood mutual information entropy is integrated into the traditional neighborhood entropy attribute reduction, and a neighborhood entropy attribute reduction algorithm based on minimizing neighborhood mutual information is proposed. The simulation results show that the algorithm can further improve the degree of independence of attributes in attribute reduction results, and has higher reduction performance than the related attribute reduction algorithms.

Key words: rough set; attribute reduction; hybrid information system; neighborhood entropy; neighborhood mutual information entropy

收稿日期: 2019—06—20; 修回日期: 2019—07—05

基金项目: 江苏高校哲学社会科学研究项目(2018SJA1927); 江苏省教育厅教改课题(2017JSJG081); 江苏高校“青蓝工程”资助项目

1 引言

属性约简^[1]也称之为特征选择,是粗糙集理论^[2]的重要研究内容,其主要目的是在不降低原始数据集分类性能的前提下,对数据集的冗余和不相关属性进行约简.由于现实环境下数据类型的多样性,基于离散型和连续型混合的信息系统是目前属性约简的重点研究对象.

基于邻域熵的各类属性约简算法^[3-12]中,它们的主要方法是通过邻域熵作为启发式函数进行属性约简的搜索^[9-12],但是邻域熵搜索到的属性未考虑属性之间的独立性,因此该属性约简结果中可能存在着冗余属性^[13].为了改进邻域熵属性约简的性能,本文提出了基于最小化邻域互信息的邻域熵属性约简算法,最后通过实验分析来证明所提算法的有效性和优越性.

2 基本理论

2.1 邻域粗糙集

在属性约简的研究中,数据集通常表示成信息系统的形式,对于一个信息系统 $IS = (U, AT)$,其中 U 为这个信息系统的论域,即对应数据集的样本集, AT 为信息系统的属性集,即对应数据集的特征集.当属性集 AT 中离散型属性和连续型属性并存时,此信息系统又称为混合型信息系统^[3].

在粗糙集理论中,邻域粗糙集模型^[3]通过建立邻域关系来对混合信息系统的论域进行粒化,从而达到对目标知识的近似逼近^[4].

定义 1 设混合信息系统 $IS = (U, AT)$, 属性集 $A \subseteq AT$, 令邻域半径为 δ , 那么属性集 A 在论域 U 上建立的邻域关系定义为

$$N_A^\delta = \{(x, y) \mid d_A(x, y) \leq \delta, \forall x, y \in U\}$$

$d_A(x, y)$ 表示的是对象 x 和 y 在属性集 A 下的广义距离度量,通常 $d_A(x, y)$ 用欧氏距离来进行计算.

根据邻域关系,可以对混合型信息系统的论域构造出一组邻域粒化.

定义 2 设混合信息系统 $IS = (U, AT)$, $U = \{x_1, x_2, \dots, x_n\}$, 属性集 $A \subseteq AT$, 令邻域半径为 δ , 属性集 A 在论域 U 上确定的邻域关系为 N_A^δ , 那么 N_A^δ 在论域 U 上诱导的邻域粒化表示为

$$U/N_A^\delta = \{\delta_A(x_1), \delta_A(x_2), \dots, \delta_A(x_n)\}$$

式中, $\delta_A(x_i), \forall x_i \in U$ 表示整个邻域粒化 U/N_A^δ 的信息粒子,定义为

$$\delta_A(x) = \{y \in U \mid (x, y) \in N_A^\delta\}$$

同时 $\delta_A(x_i), \forall x_i \in U$ 在邻域粗糙集模型中也称之为邻域粒或邻域类,是邻域粗糙集进行模型计算的基本单位.

定义 3 设混合信息系统 $IS = (U, AT)$, 令邻域半径为 δ , 属性集 $A \subseteq AT$ 在论域 U 上确定的邻域关系为 N_A^δ . 对于近似对象集 $X \subseteq U$ 关于 N_A^δ 的邻域下近似和邻域上近似分别定义为

$$\underline{apr}_A^\delta(X) = \{x \in U \mid \delta_A(x) \subseteq X\}$$

$$\overline{apr}_A^\delta(X) = \{x \in U \mid \delta_A(x) \cap X \neq \emptyset\}$$

2.2 基于邻域熵的属性约简

在混合信息系统 $IS = (U, AT)$ 中,邻域关系将信息系统的论域粒化成一组粒空间,类似于等价关系下的信息熵模型,并进一步地提出了邻域熵模型^[8],并根据邻域熵提出了混合信息系统的属性约简算法.

定义 4 设混合信息系统 $IS = (U, AT)$, $U = \{x_1, x_2, \dots, x_n\}$, 令邻域半径为 δ , 属性集 $A \subseteq AT$ 在论域 U 上确定的邻域关系为 N_A^δ . 那么论域 U 下的邻域熵定义为

$$NE_\delta(A) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i)|}{n}$$

对于 $\forall x_i \in U$, 由于 $\delta_A(x_i) \subseteq U$ 且 $\delta_A(x_i) \neq \emptyset$, 所以 $0 < \frac{|\delta_A(x_i)|}{n} \leq 1$, 则满足 $0 \leq NE_\delta(A) \leq \log n$. 其中当 $NE_\delta(A) = 0$ 时, 即 $\forall x_i \in U$ 都有 $\delta_A(x_i) = U$, 当 $NE_\delta(A) = \log n$, 即 $\forall x_i \in U$ 都有 $\delta_A(x_i) = \{x_i\}$.

在邻域熵的基础上, Hu 等学者提出了邻域联合熵.

定义 5 设混合信息系统 $IS = (U, AT)$, $U = \{x_1, x_2, \dots, x_n\}$, 令邻域半径为 δ , 属性集 $A, B \subseteq AT$ 在论域 U 上确定的邻域关系分别为 N_A^δ 和 N_B^δ . 那么论域 U 下的联合邻域熵定义为

$$JNE_\delta(A, B) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{n}$$

综合邻域熵和联合邻域熵所描述的信息不确定性,可以得到属性之间的条件邻域熵.

定义 6 设混合信息系统 $IS = (U, AT)$, $U =$

$\{x_1, x_2, \dots, x_n\}$, 令邻域半径为 δ , 属性集 $A, B \subseteq AT$ 在论域 U 上确定的邻域关系分别为 N_A^δ 和 N_B^δ . 那么论域 U 下属性集 A 关于 B 的条件邻域熵定义为

$$CNE_\delta(A | B) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{|\delta_B(x_i)|}$$

特别地, 混合型信息系统 $IS = (U, AT)$, 决策属性为 D , 那么决策属性 D 关于属性集 $A \subseteq AT$ 的条件邻域熵为

$$CNE_\delta(D | A) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|[x_i]_D \cap \delta_A(x_i)|}{|\delta_A(x_i)|}$$

式中, $[x_i]_D$ 为对象 x_i 在 D 下的等价类.

根据信息论知识, 定义 4 中所示的邻域熵表明了某个单一属性集下的信息量, 定义 5 所示的联合邻域熵表明了两个属性集之间公共的信息量, 而定义 6 所示的条件熵则表明了某个属性集在另一个属性集下的信息量, 这也反映出两个属性集在信息量视角下的依赖程度, 基于此提出了条件邻域熵的属性约简.

定义 7 设混合信息系统 $IS = (U, AT)$, 决策属性为 D , 令邻域半径为 δ . 若属性集 $R \subseteq AT$ 是整个信息系统的条件邻域熵属性约简, 那么

$$(1) CNE_\delta(D | R) = CNE_\delta(D | AT);$$

$$(2) \forall a \in R, CNE_\delta(D | R) < CNE_\delta(D | R - \{a\}).$$

在定义 7 中, 条件(1)表明决策属性 D 在约简集 R 和属性集 AT 下具有相同的条件邻域熵, 条件(2)表明约简集 R 中不存在冗余的属性, 这保证了属性约简的极小性. 根据定义 7, 混合型信息系统下基于条件邻域熵的属性约简如算法 1 所示.

算法 1: 混合型信息系统下基于条件邻域熵的属性约简算法

输入: 混合型信息系统 $IS = (U, AT)$, 决策属性 D , 邻域半径 δ .

输出: 属性约简 red

Step1. 初始化 $red = \emptyset$.

Step2. 对于属性 $\forall a \in AT - red$,

计算条件邻域熵 $CNE_\delta(D | red)$ 和 $CNE_\delta(D | red \cup \{a\})$.

计算 $sig(a) = CNE_\delta(D | red) - CNE_\delta(D | red \cup \{a\})$

Step3. 选择出 $\forall a \in AT - red$ 中 $sig(a)$ 值最大

的属性, 记为 a_{max} .

Step4. 若 $sig(a_{max}) > 0$, 则 $red = red \cup \{a_{max}\}$, 并转入 Step2. 否则, 转入 Step5.

Step5. 返回属性约简结果 red .

在混合型信息系统 $IS = (U, AT)$ 中, 若 $|U| = n$ 且 $|AT| = m$, 根据文献[8], 算法 1 的时间复杂度为 $O(m^2 \cdot n^2)$.

3 基于最小化邻域互信息的邻域熵属性约简

3.1 邻域互信息熵

在邻域粗糙集的信息熵模型中, 邻域条件熵反映了属性之间依赖程度的信息量度量, 根据香农的信息熵理论, 可以进一步地推导出邻域粗糙集模型下的邻域互信息熵.

定义 8 设混合信息系统 $IS = (U, AT)$, $U = \{x_1, x_2, \dots, x_n\}$, 令邻域半径为 δ , 属性集 $A, B \subseteq AT$ 在论域 U 上确定的邻域关系分别为 N_A^δ 和 N_B^δ . 那么论域 U 下属性集 A 和属性集 B 的邻域互信息熵定义为

$$NME_\delta(A; B) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i)| \cdot |\delta_B(x_i)|}{n \cdot |\delta_A(x_i) \cap \delta_B(x_i)|}$$

特别地, 混合信息系统 $IS = (U, AT)$, 决策属性为 D , 那么决策属性 D 关于属性集 $A \subseteq AT$ 的邻域互信息熵为

$$NME_\delta(D; A) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|[x_i]_D| \cdot |\delta_A(x_i)|}{n \cdot |[x_i]_D \cap \delta_A(x_i)|}$$

邻域互信息熵与邻域熵、邻域联合熵以及邻域条件熵满足如下关系.

定理 1 设混合信息系统 $IS = (U, AT)$, $U = \{x_1, x_2, \dots, x_n\}$, 令邻域半径为 δ , 属性集 $A, B \subseteq AT$ 在论域 U 上确定的邻域关系分别为 N_A^δ 和 N_B^δ . 那么有

$$(1) NME_\delta(A; B) = NME_\delta(B; A)$$

$$(2) NME_\delta(A; B) = NE_\delta(A) + NE_\delta(B) - JNE_\delta(A, B)$$

$$(3) NME_\delta(A; B) = NE_\delta(A) - CNE_\delta(A | B) = NE_\delta(B) - CNE_\delta(B | A)$$

证明:

(1) 根据邻域互信息熵的定义, 有

$$NME_\delta(A; B)$$

$$\begin{aligned}
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i)| \cdot |\delta_B(x_i)|}{n \cdot |\delta_A(x_i) \cap \delta_B(x_i)|} \\
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_B(x_i)| \cdot |\delta_A(x_i)|}{n \cdot |\delta_B(x_i) \cap \delta_A(x_i)|} \\
&= NME_\delta(B; A)
\end{aligned}$$

因此(1)成立.

(2) 根据邻域熵和邻域联合熵的定义,有

$$\begin{aligned}
&NE_\delta(A) + NE_\delta(B) - JNE_\delta(A, B) \\
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i)|}{n} - \frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_B(x_i)|}{n} \\
&\quad - \left(-\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{n} \right) \\
&= -\frac{1}{n} \sum_{i=1}^n \left(\log \frac{|\delta_A(x_i)|}{n} + \log \frac{|\delta_B(x_i)|}{n} \right. \\
&\quad \left. - \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{n} \right) \\
&= -\frac{1}{n} \sum_{i=1}^n \left(\log \frac{|\delta_A(x_i)|}{n} \cdot \frac{|\delta_B(x_i)|}{n} \right. \\
&\quad \left. \cdot \frac{n}{|\delta_A(x_i) \cap \delta_B(x_i)|} \right) \\
&= -\frac{1}{n} \sum_{i=1}^n \left(\log \frac{|\delta_A(x_i)| \cdot |\delta_B(x_i)|}{n \cdot |\delta_A(x_i) \cap \delta_B(x_i)|} \right) \\
&= NME_\delta(A; B)
\end{aligned}$$

因此(2)成立.

(3) 根据邻域熵和条件邻域熵的定义,有

$$\begin{aligned}
&NE_\delta(A) - CNE_\delta(A | B) \\
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i)|}{n} \\
&\quad - \left(-\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{|\delta_B(x_i)|} \right) \\
&= -\frac{1}{n} \sum_{i=1}^n \left(\log \frac{|\delta_A(x_i)|}{n} - \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{|\delta_B(x_i)|} \right) \\
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i)|}{n} \cdot \frac{|\delta_B(x_i)|}{|\delta_A(x_i) \cap \delta_B(x_i)|} \\
&= NME_\delta(A; B)
\end{aligned}$$

同时

$$NME_\delta(A; B)$$

$$\begin{aligned}
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_B(x_i)|}{n} \cdot \frac{|\delta_A(x_i)|}{|\delta_A(x_i) \cap \delta_B(x_i)|} \\
&= -\frac{1}{n} \sum_{i=1}^n \left(\log \frac{|\delta_B(x_i)|}{n} - \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{|\delta_A(x_i)|} \right) \\
&= -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_B(x_i)|}{n} \\
&\quad - \left(-\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{|\delta_A(x_i)|} \right)
\end{aligned}$$

由于

$$NE_\delta(B) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_B(x_i)|}{n};$$

$$CNE_\delta(B | A) = -\frac{1}{n} \sum_{i=1}^n \log \frac{|\delta_A(x_i) \cap \delta_B(x_i)|}{|\delta_A(x_i)|}.$$

所以 $NME_\delta(A; B) = NE_\delta(B) - CNE_\delta(B | A)$,

综上可以得到

$$\begin{aligned}
NME_\delta(A; B) &= NE_\delta(A) - CNE_\delta(A | B) \\
&= NE_\delta(B) - CNE_\delta(B | A)
\end{aligned}$$

定理1证毕.

定理1中的(1)式表明,信息系统属性集之间的邻域互信息熵之间满足对称性,即属性集A与属性集B之间的邻域互信息熵即为属性集B与属性集A之间的邻域互信息熵.定理1中的(2)式是邻域互信息熵的一个很重要的性质,它揭示了邻域互信息熵的本质,即属性集A和属性集B之间的邻域互信息熵为属性集A和B邻域熵总和的基础上去除公共邻域熵的部分,也就是说邻域互信息熵反映了属性集A和B之间的独立程度,当邻域互信息熵越小,说明属性集A和B之间的独立性越强,反之则说明越弱.图1展示的是各类熵度量之间的关系.

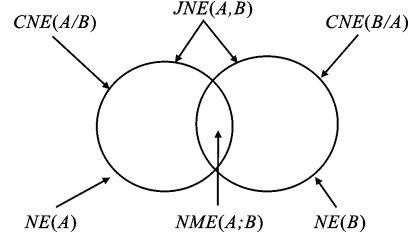


图1 邻域各类熵之间的关系

3.2 最小化邻域互信息的邻域熵属性约简

在邻域熵基础上推导出的邻域条件熵是目前混合型信息系统的一种重要的属性约简方法,并且该方法已进行了大量的改进和应用.基于邻域熵的属性约简通过邻域条件熵作为启发式函数对属性进行搜索和选择,然而对于每次选择到的属性,未考虑到该属性与已选择属性之间的邻域互信息,即选择到的属性可能与已选择的属性存在依赖关系,从而最终的属性约简结果存在冗余的属性.因此本文将基于邻域熵的属性约简算法进行改进,利用邻域互信息熵进行属性之间独立性的评估,提出一种最小邻域互信息的邻域熵属性约简算法.

在算法1的Step2中,通过属性重要度 $\text{sig}(a) = CNE_\delta(D | \text{red}) - CNE_\delta(D | \text{red} \cup \{a\})$ 进行属性的

选择,即每次选择属性 $\text{argmax}(\text{sig}(a))$ 作为约简集的候选属性. 为了将邻域互信息融入属性的评价准则中,我们定义了一种新的属性集评估函数.

定义 9 设混合信息系统 $IS = (U, AT)$, 令邻域半径为 δ , 决策属性集为 D , 属性集 $A \subseteq AT$. 定义属性 $a \in AT - A$ 在决策属性集 D 下关于 A 的属性评估为

$$F_A(a) = CNE(D \mid A \cup \{a\}) + HME(A; a)$$

由于条件邻域熵的值随属性的增加是单调递减的,即属性全集下,条件邻域熵的值最小,因此算法 1 在进行属性搜索时,选择的是使条件邻域熵值降低最快的属性. 在定义 9 中,新的评估函数由条件邻域熵和邻域互信息熵两部分组成,对于 $a \in AT - A$, 当 $F_A(a)$ 值最小时,即条件邻域熵 $CNE(D \mid A \cup \{a\})$ 和邻域互信息熵 $HME(A; a)$ 整体具有最小的结果,那么在每次搜索属性时,这样既保证了条件邻域熵快速地减少,也满足了新选择的属性具有较小的依赖性. 根据定义 9 的属性评估函数,我们提出一种改进的属性约简算法,具体如算法 2 所示.

算法 2:混合型信息系统下最小化邻域互信息的邻域熵属性约简算法

输入:混合型信息系统 $IS = (U, AT)$, 决策属性 D , 邻域半径 δ .

输出:属性约简 red .

Step1. 初始化 $\text{red} = \emptyset$.

Step2. 对于 $\forall a \in AT - \text{red}$, 计算每个属性的评估函数:

$$F_{\text{red}}(a) = CNE(D \mid \text{red} \cup \{a\}) + HME(\text{red}; a)$$

选择出 $\forall a \in AT - \text{red}$ 中 $F_{\text{red}}(a)$ 值最小的属性 $a_{\min}$, 即 $a_{\min} = \arg \min_{a \in AT - \text{red}} \{F_{\text{red}}(a)\}$.

Step3. $\text{red} = \text{red} \cup \{a_{\min}\}$, 若

$$CNE_{\delta}(D \mid \text{red}) = CNE_{\delta}(D \mid AT)$$

那么跳转入 Step4, 否则跳转入 Step2.

Step4. 返回属性约简结果 red .

在混合信息系统 $IS = (U, AT)$ 中, 设 $|U| = n$ 且 $|AT| = m$, 算法 2 的时间复杂度与算法 1 相同, 同样为 $O(m^2 \cdot n^2)$.

4 实验分析

实验中将本文所提出的算法与其他同类型的算

法对同一组数据集进行属性约简,然后比较属性约简的结果,从而比较各类算法之间的优越性. 表 1 所示的是实验中的数据集,这些数据集均选择于 UCI 机器学习数据集,并且所有的数据集均为连续型属性和离散型属性混合的数据集,在实验前所有连续型属性均标准化至 $[0, 1]$ 区间. 实验中所有的算法均采用 Matlab2012 进行编程运行,所运行的计算机配置为 Intel Core i5 7500 3.4 GHz CPU 和 DDR4 8 GB 内存.

表 1 实验数据集

序号	数据集名称	对象	属性
1	Wdbc	569	31
2	Mess	1 151	19
3	German	1 000	19
4	Sick	2 800	28
5	Musk	6 598	166
6	Gearbox	1 603	72

4.1 实验设置

本实验中,所选取的对比算法分别为基于基于邻域粗糙集正区域的属性约简算法^[3](记为算法 A),基于粒化单调的条件邻域熵的属性约简算法^[9](记为算法 B),基于模糊粗糙集信息熵的属性约简算法^[14](记为算法 C),基于邻域熵的属性约简算法^[8](记为算法 D). 在本文所提出的算法中,其中包含一个输入参数,即邻域半径 δ , 它的取值不同将对最终的属性约简结果产生一定的影响. 本实验采用已有的研究成果^[3-4],即选择 $\delta=0.1$ 进行实验. 实验中将通过比较各个算法的属性约简集、属性约简分类精度以及属性约简效率来验证算法的优越性.

4.2 属性约简的结果比较

表 2 是所有算法在每个数据集下的属性约简结果,为约简集中属性的数量. 观察可得算法 A 在大部分数据集下约简得到的属性较多,例如数据集 Wdbc, German, Sick 和 Musk. 这主要是由于该算法是基于邻域粗糙集的正区域进行约简,其效果要劣于信息熵的方法. 本文所提出的算法在大部分数据集中拥有较少属性约简结果,例如数据集 Wdbc, 本文算法约简出了 8 个属性,而算法 A 的约简出的属性个数为 12, 算法 B 的约简出的属性个数为 9, 例如数据集 Gearbox, 本文算法约简出了 13 个属性,而算法 B 的约简出的属性个数为 19, 算法 C 的约简出的属性个数为 15. 因此可以看出,本文所提出的属性约简算法能够约简出更少的属性约简结果,这

主要是由于本文算法通过邻域互信息的度量对选择出的属性进行进一步地过滤,保证了每次选择的属性具有较高的独立性,而其他的属性约简算法仅通过对应启发式函数进行属性的搜索,因此选择出的属性会更多一些.

表 2 属性约简结果

数据集名称	原始属性集	算法 A	算法 B	算法 C	算法 D	本文算法
Wdbc	31	12	9	11	10	8
Mess	19	8	9	10	10	7
German	19	9	8	7	8	7
Sick	28	13	12	12	10	9
Musk	166	24	22	17	19	20
Gearbox	72	18	19	15	16	13

4.3 属性约简结果的分类精度比较

对于各种属性约简算法得到的约简结果,本小节通过支持向量机(SVM)分类算法和决策树(DT)分类算法对属性约简结果进行十折交叉分类训练和分类预测,最终会得到每个约简结果在对应数据集下的分类精度,具体如图 2 和图 3 所示.

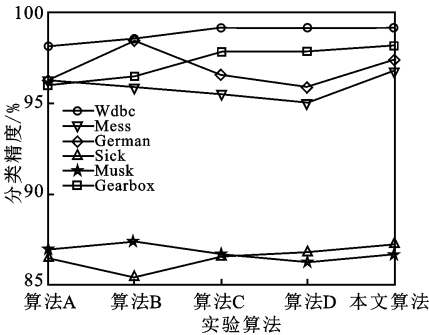


图 2 SVM 分类精度结果

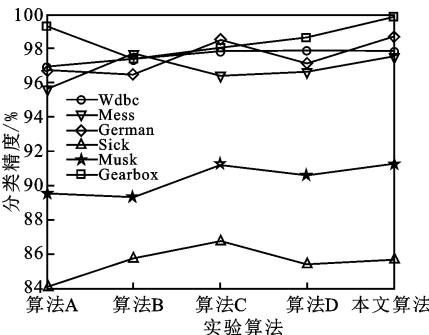


图 3 DT 分类精度结果

在图 2 所示的 SVM 分类精度中,观察可以发现 5 种算法的分类精度差距不是特别大,例如数据集 Wdbc,最高的分类精度是本文算法得到的结果,

其值约为 99%,最低为算法 A,其值约为 97%.数据集 Musk,最高的分类精度是算法 B 得到的结果,其值约为 87%,最低为算法 D 的结果,其值为 86%.本文所提出的属性约简算法,在 Wdbc, Mess, Sick 和 Gearbox 数据集下拥有较高的分类精度.观察图 3 的 DT 分类精度结果,也能够得到同样的结论.综合分析表明,本文算法选择出的属性约简结果同样是有有效的.

4.4 属性约简结果的处理效率比较

本实验将所有算法对每个数据集进行重复属性约简 10 次,然后将 10 次约简的运行时间取平均值作为对应算法的最终用时结果,如图 4 所示.观察图 4 可以发现算法 D 在所有数据集中拥有最少的用时结果,本文算法的用时结果仅次于算法 D,而算法 A 的用时结果最大.这主要是由于算法 D 基于邻域熵进行属性约简,只需计算每个对象的邻域类便可以快速地完成任务,本文算法是在邻域熵属性约简的基础上进行的改进,其中需要对邻域熵选择出的属性再进行邻域互信息的度量,这消耗了一定的计算时间.因此本文的算法同样具有较高的属性约简效率.

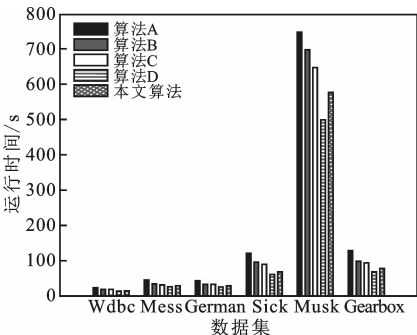


图 4 各个算法运行时间比较

5 结束语

由于邻域熵在进行属性约简时未考虑属性之间的独立性,本文提出一种基于最小邻域互信息的邻域熵属性约简算法.在邻域熵的基础上提出了邻域互信息熵的定义,并证明相关性质,理论分析表明邻域互信息熵可以作为混合型信息系统属性之间的独立性度量,最后将其融入传统的邻域熵属性约简算法中,提出了一种改进的属性约简算法.实验分析表明了所提出算法的有效性.在将来的研究中,将进一步探索动态数据环境下最小化邻域互信息的属性约简问题.

参考文献:

- [1] 邬阳阳, 汤建国. 大数据背景下粗糙集属性约简研究进展[J]. 计算机工程与应用, 2019, 55(6):31-38.
- [2] PAWLAK Z. Rough sets[J]. International Journal Computer Information Science, 1982, 11 (5): 341-356.
- [3] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18):3577-3594.
- [4] HU Q H, YU D R, XIE Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34 (2):866-876.
- [5] LIU Y, HUANG W L, JIANG Y L, et al. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences, 2014(271):65-81.
- [6] 姚晟, 汪杰, 徐风, 等. 不一致邻域粗糙集的不确定性度量和属性约简[J]. 小型微型计算机系统, 2018, 39 (4):700-706.
- [7] 徐波, 张贤勇, 冯山. 邻域粗糙集的加权依赖度及其启发式约简算法[J]. 模式识别与人工智能, 2018(3): 256-264.
- [8] HU Q H, ZHANG L, ZHANG DAVID, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert Systems with Applications, 2011, 38 (9): 10737-10750.
- [9] 周艳红, 张贤勇, 莫智文. 粒化单调的条件邻域熵及其相关属性约简[J]. 计算机研究与发展, 2018, 55 (11):55-65.
- [10] ZHAO H, QIN K Y. Mixed feature selection in incomplete decision table [J]. Knowledge-Based Systems, 2014(57):181-190.
- [11] 姚晟, 徐风, 赵鹏, 等. 基于邻域量化容差关系粗糙集模型的特征选择算法[J]. 模式识别与人工智能, 2017(5):34-46.
- [12] 陈迎春, 李鸥, 孙昱. 基于聚类离散化和变精度邻域熵的属性约简[J]. 控制与决策, 2018, 33(8):66-73.
- [13] ESTÉVEZ P, MEMBER S, PEREZ C, et al. Normalized mutual information feature selection [J]. IEEE Transactions on Neural Networks, 2009, 20 (2): 189-201.
- [14] 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):1-15.

作者简介:

刘正男, (1974-), 硕士, 副教授. 研究方向为信息与信号处理和人工智能. E-mail: furh1978@163.com.

陈雪勤女, (1970-), 博士, 副教授. 研究方向为深度学习.

张书锋男, (1980-), 硕士, 高级工程师. 研究方向为计算机软件设计和开发及移动互联网.