

基于模糊粗糙集依赖度的两步属性约简方法

王世强¹, 张登福¹, 毕笃彦¹, 张立东²

(1. 空军工程大学 工程学院, 西安 710038; 2. 空军装备研究院 雷达与电子对抗研究所, 北京 100085)

摘 要: 为获取连续属性值数据集的最小属性子集, 提出了一种两步约简方法. 该方法以模糊粗糙集模型为基础, 将描述条件属性和决策属性依赖关系的模糊依赖度概念进行了扩展, 使其能对条件属性之间的依赖关系进行度量. 利用属性与类别之间的依赖度选出候选属性集, 然后根据单个属性与类别和属性之间的依赖度对候选属性集进行约简. 仿真结果表明, 该方法在有效降低属性维数的同时一定程度上保证了分类正确率.

关键词: 属性约简; 粗糙集理论; 模糊粗糙集; 模糊等价关系; 隶属函数; 依赖度

中图分类号: TP 181

文献标志码: A

文章编号: 0254-0037(2013)06-0828-07

Two-step Attribute Reduction Method Based on Fuzzy Rough Sets Dependency

WANG Shi-qiang¹, ZHANG Deng-fu¹, BI Du-yan¹, ZHANG Li-dong²

(1. Institute of Engineering, Air Force Engineering University, Xi'an 710038, China;

2. Rader and Electric Countermeasure Research Institute, Air Force Equipment Academy, Beijing 100085, China)

Abstract: To acquire the minimum attribute reduction of the dataset with continual attribute value, a two-step reduction method is proposed based on fuzzy rough sets. The concept of dependency is extended and the dependency relationship between condition attributes can be measured on the basis of the extended concept. The candidate attributes are first selected by calculating the attribute importance. Then, the candidate attributes are reduced using the dependency between attributes and of the single attribute importance. The redundant attribute is reduced via this operation. Experimental results show that the strategy can efficiently reduce the attribute dimension without sacrificing the classification accuracy.

Key words: attribute reduction; rough set theory; fuzzy rough sets; fuzzy equivalence relationship; membership function; dependency

粗糙集理论是 Pawlak 提出的一种能处理不精确和不完整信息的数学工具^[1-2]. 目前, 该理论已被成功地应用于机器学习、知识发现、数据挖掘、决策分析等诸多领域. 其中, 以不降低信息系统中分辨不同对象能力的属性约简问题一直是粗糙集理论研究的核心问题之一^[3]. 然而, 经典粗糙集理论以分明等价关系为基础, 要求不同对象的属性值之间具有精确的等价关系才能构成等价类, 主要用来处理

清晰、有限且离散的属性值. 对于本身具有连续属性值的数据集却不能直接进行处理, 而是需要以不同方法对属性值进行离散化, 如云变换离散化方法^[4]、监督型局部离散化方法^[5]、基于属性重要性的离散化方法^[6]等. 但是离散处理不能很好地保留属性值在实数值上存在的差异, 使得信息内容的完整性难以保证, 从而导致不同程度的信息损失, 因此利用粗糙集理论处理这类问题具有一定的局限性.

收稿日期: 2011-12-28.

基金项目: 国家自然科学基金资助项目(61175029); 航空科学基金资助项目(20101996009).

作者简介: 王世强(1982—), 男, 博士研究生, 主要从事智能信息处理、模式识别等方面的研究. E-mail: wunsicon@163.com.

针对这一问题,研究人员提出了联合利用粗糙集和模糊集^[7-11]来解决模糊属性信息表中的属性约简问题,其中,基于模糊粗糙集的属性约简在近几年备受关注,并取得了一定的研究成果^[12-15].文献[12]定义了模糊点和等价类基元的概念,并据此给出一种模糊下近似的定义,进而利用区分矩阵完成属性约简;文献[13]给出了一种基于属性集满意度的约简方法;文献[14-15]提出用依赖度衡量属性子集的重要性,并将依赖度函数作为启发式信息函数的约简方法,该方法对于属性数目较少的数据集具有较好的约简效果,但许多实际分析中该方法不收敛^[16].

本文在前人工作启发下,对模糊粗糙集中的依赖度概念进行了扩展,使其能对条件属性之间的依赖关系进行度量,根据条件属性间的依赖度给出了候选属性集和冗余属性集的定义及基于该依赖度的约简算法.此算法综合度量属性与类别和属性之间的依赖度,通过属性间的依赖关系再次确定属性的重要性,进一步剔除约简子集中包含的冗余属性,从而得到最小属性集. UCI 测试数据集试验结果表明了该方法的有效性.

1 模糊粗糙集模型

在经典粗糙集理论中,将近似对象从分明集扩展到模糊集,并用论域上的模糊等价关系代替分明等价关系即可得到模糊粗糙集.经典粗糙集研究对象是分明等价类,而模糊粗糙集研究对象是模糊等价类,是将论域上的对象在模糊等价关系下划分成为模糊等价类.

设 $I = (U, A \cup Q)$ 为一个信息系统,其中 U 表示对象的非空有限集,即论域; A 表示条件属性的非空有限集,在不指明的情况下,属性是指条件属性; Q 表示决策属性的非空有限集,即类别属性集.属性集 $P \subseteq A$ 对应一个不可分辨的等价关系,在不引起歧义的情况下,将此等价关系简记为 P ,将以此信息系统为前提展开研究.

1.1 模糊等价类

以下记论域 U 上的模糊等价关系为 P ,对象 x 和 y 之间的相似度为 $u_P(x, y) = u_P(y, x)$,满足三角模算子的条件,即^[17]

- 1) 自反性: $\forall x \in U, u_P(x, x) = 1$;
- 2) 对称性: $\forall x, y \in U, u_P(x, y) = u_P(y, x)$;
- 3) 传递性: $\forall x, y, z \in U, \min(u_P(x, y), u_P(y, z)) \leq u_P(x, z)$.

在模糊等价关系 P 概念基础上,即可定义对象的模糊等价类.

定义1 设 P 是论域 U 上的一个模糊相似关系,定义对象 $x \in U$ 的模糊等价类 $[x]_P$ 如下^[17]:

$$\mu_{[x]_P}(y) = \mu_P(x, y), y \in U \quad (1)$$

当 P 是 U 上的普通等价关系时,式(1)定义了 P 的等价类 $[x]_P$.一般情况下,式(1)表示论域 U 中与对象 x 邻近的所有对象的集合,它是论域上的一个模糊集,与分明粗糙集定义的等价类具有相同的性质,模糊等价类 $[x]_P$ 中的任意2个对象也是不可分辨的.

1.2 模糊粗糙集的上下近似

设 F_i 是属于 U/P 的模糊等价类, $U/P = F = \{F_1, F_2, \dots, F_n\}$ 表示论域 U 上的一个模糊划分,属性子集 $P \subseteq A$, X 表示 U 上的任意模糊集合,则可通过模糊条件集 F 来近似 X ,即用 F 中的子集 F_i 包含在 X 中的可能度与必然度描述 F 对 X 的近似程度,这种描述称为在关系 P 下 X 的模糊上下近似,定义如下^[18]:

$$\begin{aligned} \mu_{\bar{P}X}(F_i) &= \sup_x u_{F_i \cap X}(x) = \\ &\sup_x \min\{u_{F_i}(x), \mu_X(x)\} \quad \forall i \end{aligned} \quad (2)$$

$$\begin{aligned} \mu_{\underline{P}X}(F_i) &= \inf_x u_{(1-F_i) \cap X}(x) = \\ &\inf_x \max\{1 - u_{F_i}(x), \mu_X(x)\} \quad \forall i \end{aligned} \quad (3)$$

式中: $u_{F_i}(x)$ 表示 U 中对象 x 包含在 F_i 中的程度; $u_X(x)$ 表示 x 包含在 X 中的程度; $\mu_{\bar{P}X}(F_i)$ 表示 F_i 包含在 X 中的必然度,即 F_i 必然包含在 X 中的程度.

在关系 P 下通过子集 F_i 描述 X 时, x 必然包含在 X 中的程度 $\mu_{\bar{P}X}(x^{F_i})$ 即为 x 包含在 F_i 与 X 交集的程度,表示为

$$\begin{aligned} \mu_{\bar{P}X}(x^{F_i}) &= \mu_{F_i \cap \underline{P}X}(x) = \\ &\min(\mu_{F_i}(x), \mu_{\underline{P}X}(F_i)) \end{aligned} \quad (4)$$

由于 U/P 划分的模糊等价类互不相交,因此 x 必然包含在 X 中的程度 $\mu_{\bar{P}X}(x)$ 为所有模糊条件子集 F_i 描述模糊集合 X 时最大的 $\mu_{\bar{P}X}(x^{F_i})$,考虑到有限论域 U 不能推广到一般情况下,从而有

$$\begin{aligned} \mu_{\bar{P}X}(x) &= \sup_{F_i \in U/P} \mu_{\bar{P}X}(x^{F_i}) = \\ &\sup_{F_i \in U/P} \min(\mu_{F_i}(x), \mu_{\underline{P}X}(F_i)) \end{aligned} \quad (5)$$

因此,模糊下近似可重新定义为

$$\begin{aligned} \mu_{\underline{P}X}(x) &= \sup_{F_i \in U/P} \min(\mu_{F_i}(x), \\ &\inf_{y \in U} \max\{1 - \mu_{F_i}(y), \mu_X(y)\}) \end{aligned} \quad (6)$$

同样可定义模糊上近似 $\mu_{\bar{P}X}$,称序对 $\langle \bar{P}X, \underline{P}X \rangle$ 为模糊集 X 在 U 上的模糊粗糙集.模糊粗糙集模型不

仅将近似对象从分明集扩展到了模糊集,而且将论域上的等价关系扩展到了模糊相似关系.

1.3 模糊等价类隶属函数的确定

论域 U 中,对象 x 可能属于不止一个模糊属性决定的模糊等价类,因此必须考虑由多模糊属性决定的模糊等价类,即由单个模糊属性决定的模糊等价类的笛卡儿乘积.

考察分明粗糙集 U/P 决定了一个包含由所有不可分辨对象组成不同集合的簇集. 设论域 $U = \{x_1, x_2, \dots, x_n\}$, 属性集 P 包含 m 个属性 $\{a_1, a_2, \dots, a_m\}$, 对于集合 B 和 C 定义运算符 $\otimes$:

$$B \otimes C = \{X \cap Y: \forall X \in B, \forall Y \in C, X \cap Y \neq \emptyset\} \quad (7)$$

则 U/P 可通过式(8)计算:

$$U/P = \otimes \{a_i \in P: U/IND(a_i), i=1, 2, \dots, m\} = U/\{a_1\} \otimes U/\{a_2\} \otimes \dots \otimes U/\{a_m\} \quad (8)$$

考虑 $m=2$ 时的情形,

$$\begin{aligned} U/P &= U/\{a_1\} \otimes U/\{a_2\} = \\ &= \{X_1^{a_1}, X_2^{a_1}, \dots\} \otimes \{X_1^{a_2}, X_2^{a_2}, \dots\} = \\ &= \{X_1^{a_1} \cap X_1^{a_2}, X_1^{a_1} \cap X_2^{a_2}, \dots, \\ &X_2^{a_1} \cap X_1^{a_2}, X_2^{a_1} \cap X_2^{a_2}, \dots\} \end{aligned} \quad (9)$$

式中:集合 $\{X_1^{a_1}, X_2^{a_1}, \dots\}$ 和 $\{X_1^{a_2}, X_2^{a_2}, \dots\}$ 分别表示由属性 $\{a_1\}$ 和 $\{a_2\}$ 决定的 U 的划分. 分明粗糙集中,分明集 $X_j^{a_i} = [x_j]_{\{a_i\}}$. 记由 P 确定的 U 的划分为 $U/P = \{X_1, X_2, \dots, X_n\}$, $\forall X_i \in X$ 都是分明集,用隶属函数表示 $x \in U$ 与 X_i 的关系,则有

$$\mu_{X_i}(x) = \begin{cases} 1 & x \in X_i \\ 0 & x \notin X_i \end{cases} \quad (10)$$

在根据实际需要连续属性区间进行划分以确定各模糊属性的模糊等价类时,模糊条件属性被划分出的每个子区间对应于该属性的一个模糊等价类.

模糊粗糙集中,单个属性 $\{a_i\}$ 仍然对应一个不可分辨关系的等价关系,仍简记为 $\{a_i\}$. 则模糊粗糙集中 U/P 的计算与分明粗糙集类似,区别在于:集合 $\{X_1^{a_1}, X_2^{a_1}, \dots\}$ 和 $\{X_1^{a_2}, X_2^{a_2}, \dots\}$ 分别表示由属性 $\{a_1\}$ 和 $\{a_2\}$ 决定的 U 的模糊划分, $X_j^{a_i}$ 表示由某种属性区间分割算法对属性 a_i 进行分割时得到的第 j 个区间,在此区间内所有的对象都不可分辨,为了与分明粗糙集进行区分,令 $F_j^{a_i} = X_j^{a_i}$. 记由 P 确定的 U 的模糊划分为 $U/P = \{F_1, F_2, \dots, F_n\}$, 由于 $\forall F_i \in U/P$ 都是模糊集,因此对象 x_k 隶属于 F_i 的程度可通过计算对象 x_k 隶属于形成 F_i 的所有模糊集交集的程度

来获取.

同样考虑 $m=2$ 时的情形,设 $F_i = F_r^{a_g} \cap F_t^{a_h}$, 其中 $g, h, r, t=1$ 或 2 , 则用隶属函数表示 x 与 F_i 的关系时,有

$$\mu_{F_i}(x) = \mu_{F_r^{a_g} \cap F_t^{a_h}}(x) = \min(\mu_{F_r^{a_g}}(x), \mu_{F_t^{a_h}}(x)) \quad (11)$$

当 $m>2$ 时, x 和 F_i 的关系与 $m=2$ 时得到的隶属函数类似,此处不再赘述.

1.4 模糊依赖性分析

经典粗糙集理论中的分明正域定义为下近似的并集. 将此概念扩展到模糊正域,并结合式(6)和式(11),得到对象 x 关于模糊正域的隶属度:

$$\mu_{\text{POS}_P(Q)}(x) = \sup_{x \in U/Q} \mu_{P_X}(x) \quad (12)$$

根据隶属度的含义以及模糊正域的定义,可求出模糊粗糙集条件下决策属性 Q 对条件属性集 P 的依赖性^[14]:

$$k' = \gamma'_P(Q) = |\mu_{\text{POS}_P(Q)}(x)| / |U| = \sum_{x \in U} \mu_{\text{POS}_P(Q)}(x) / |U| \quad (13)$$

式中 $|U|$ 表示 U 的势,即 U 中对象的个数. 与经典粗糙集依赖性类似,式(13)表征了仅用 P 中的信息时可分辨对象在整个数据集中所占比例,依据式(13),即为 $\mu_{\text{POS}_P(Q)}$ 的势除以论域 U 的势. 式(13)同样表征了 Q 依赖于 P 的程度, P 以比例 k' 决定 Q 的取值,即决定分类效果, k' 越大,以属性 P 为依据进行分类时所取得效果越好.

式(13)所描述的依赖性可推广到属性之间依赖性,取 P 和 Q 都为单属性集合,即 $P = a_i, Q = a_j, \forall a_i, a_j \in A$, 则

$$k' = \gamma'_{a_i a_j} = \gamma'_{a_i}(a_j) = |\mu_{\text{POS}_{a_i a_j}}(x)| / |U| = \sum_{x \in U} \mu_{\text{POS}_{a_i a_j}}(x) / |U| \quad (14)$$

式(14)表征了属性 a_j 依赖于属性 a_i 的程度, a_i 以比例 k' 决定 a_j 的取值, k' 越大, a_i 决定 a_j 取值的能力越强.

2 基于模糊依赖度的属性约简

基于模糊粗糙集依赖度的属性约简算法包括 2 个步骤:一是确定候选属性集;二是确定候选属性集中是否存在冗余属性. 结合以上分析,给出了 2 个定义,则基于模糊依赖度的属性约简主要由这 2 个定义来完成.

2.1 约简方法

定义 2 设 $a_j \in A$, 记 $a_i = \max\{a_j \mid \gamma'_{a_j}(Q) = 1\}$,

$2, \dots, m\}$; 计算 $\gamma_{a_j}'(Q)$ ($j \neq i$) 对于给定阈值 T_h , 如果 $\gamma_{a_j}'(Q) > T_h$, 则称 a_j 为第一候选属性, 如果 $\gamma_{a_j}'(Q) \leq T_h$ 且 $\gamma_{\{a_i, a_j\}}'(Q) \leq \gamma_{a_i}'(Q)$ 则称 a_j 为冗余属性; 如果 $\gamma_{a_j}'(Q) \leq T_h$ 且 $\gamma_{\{a_i, a_j\}}'(Q) < \gamma_{a_i}'(Q)$ 则称 a_j 为第二候选属性. 第一和第二候选属性组成的集合称为候选属性集.

定义 3 设确定的候选属性集为 C , 给定 $a_i \in C$, 对于 $\forall a_j \in C$ ($i \neq j$), 如果 $\gamma_{a_i, a_j}' = 1$, $\gamma_{a_j, a_i}' = 0$, 或者 $\gamma_{a_i, a_j}' \approx 1$, $\gamma_{a_j, a_i}' \approx 0$, 那么称属性 a_j 是第一冗余属性, 否则称 a_j 为从属属性; 对于从属属性 a_j , 如果 $\gamma_{a_j}'(Q) \leq \gamma_{a_i}'(Q)$ 且 $\gamma_{a_j, a_i}' > \gamma_{a_i, a_j}'$ 则 a_j 为非冗余属性, 如果 $\gamma_{a_j}'(Q) \leq \gamma_{a_i}'(Q)$ 且 $\gamma_{a_j, a_i}' \leq \gamma_{a_i, a_j}'$ 则称 a_j 为第二冗余属性. 第一和第二冗余属性组成的集合称为冗余属性集.

根据模糊依赖性的分析及定义 2 和定义 3, 提出一种基于模糊依赖度的两步属性约简 (two-step attribute reduction based on fuzzy dependency, TARFD) 算法. 该算法以候选属性集为起点, 依据定义 3 得到的冗余属性, 逐次剔除重要度较小的属性, 直到满足终止条件为止.

算法: TARFD

步骤 1 令 $T_h = 0$, 由定义 2 确定候选属性集 C , 对属性 $a_i \in C$, 计算决策属性 Q 对 a_i 依赖度 $\gamma_{a_i}'(Q)$;

步骤 2 将 $\{\gamma_{a_i}'(Q) \mid i = 1, 2, \dots, m, m = |C|\}$ 按照降序排列, 得到一个属性集合, 记为 T , $T = \{\gamma_{a_{i1}}'(Q) \mid \gamma_{a_{i1}}'(Q) \geq \gamma_{a_{i2}}'(Q) \cdots \geq \gamma_{a_{il}}'(Q) \cdots \geq \gamma_{a_{im}}'(Q)\}$, 其中 $\{i1, i2, \dots, il, \dots, im\}$ 为 i 的一排列;

步骤 3 令 $k = 1$;

步骤 4 令 $j = 1$, 如果 $j \geq m - k$, 则终止算法并转到步骤 7;

步骤 5 计算属性间依赖度, 若 $\gamma_{a_{ij}}'(a_{i(j+k)}) \geq \gamma_{a_{i(j+k)}}'(a_{ij})$, 则剔除属性 $a_{i(j+k)}$, 即 $T = T \setminus \{a_{i(j+k)}\}$;

步骤 6 令 $j = j + 1$, 如果 $j < m - k$, 转到步骤 5, 否则令 $k = k + 1$, 转到步骤 4;

步骤 7 最后得到的 T 就是条件属性 P 相对于 Q 的一个相对约简.

2.2 复杂度分析

正如引言部分所指出的, Jensen 等根据模糊粗糙集依赖度提出了一种 QuickReduct 算法^[14]进行属性约简, 并在数据集维数较小情况下取得了较好的约简效果, 但是, 该算法的计算复杂度随着条件属性集的增加急剧增加. 假设属性集 P 包含 m 个

属性, 每个属性决定的 U 的划分包含 n 个模糊类, 则由属性集 P 决定的 U 的划分包含 n^m 个模糊类, 则最坏情况下决策属性属性集 P 的依赖度的近似计算复杂度可表示为 $O((n^m n)) = O(N(n^{m+1}))$. 最坏情况是指直到计算到最后一个属性时算法才终止.

例如, 设 $P = \{a, b\}$, 每个属性决定的划分包含 3 个模糊类, 则 Q 关于 P 依赖度需要计算 9 个模糊类. 实际上, 属性集的个数大于 10 的情况很正常, 因此再设 $|P| = 15$, 则 Q 关于 P 依赖度需要计算 3^{15} 个模糊类. 可见, 该算法具有相当大的计算复杂度.

对 TARFD 算法而言, 仅需要计算属性集合中的单属性依赖度和任意 2 个属性组合的依赖度, 即最坏情况下属性集 P 的依赖度的复杂度可表示为 $O(2n^2 C_m^2 + n C_m^1) \approx O(2n^2 C_m^2)$. 假如设 $|P| = 15$, 则 Q 关于 P 依赖度需要计算 1 935 个模糊类, 那么相对 QuickReduct 算法, TARFD 算法在最坏情况下执行效率提高了近 $3^{15}/1\,935 = 7\,415$ 倍, 说明该算法是高效的.

3 仿真结果分析

为了考察 TARFD 算法的有效性, 在 UCI 机器学习数据库^[19]上取 7 个连续值属性数据集进行实验. 数据集的基本情况见表 1.

表 1 试验 UCI 数据集基本信息
Table 1 Information of data sets from UCI

数据集	对象数	属性数	类别数
Glass	214	9	6
Iris	150	4	3
Ion	351	34	2
Isolet5	1559	617	26
Vehicle	946	18	4
Wav	1 000	21	3
Water2	527	38	3

在仿真实验前, 需要对表 1 中的数据集进行预处理, 即对数据集的属性进行模糊化. 为简明起见, 选取标准模糊化技术得到由单属性确定的模糊等价类, 并结合梯形函数和三角函数来确定各数据集中对象关于单属性的隶属度, 对象关于多模糊属性的隶属度由式 (11) 计算.

鉴于评估属性约简算法性能的指标主要体现在属性子集的大小和分类准确率上, 采用 Weka 平台

分别对选取的数据集进行属性约简,并使用 10-折交叉验证法得到 JRip^[20]和 C4.5 决策树^[21]的分类准确率. 为了进一步说明 TARFD 算法的有效性,选取 Relief-F (Re)^[14]、Information Gain (IG)^[14]、Fast Correlation-based Filter(FCBF)^[22]、OneR(OR)^[14]和 Fuzzy Discernibility Matrix-based (FDM)^[13]5 种不同的属性约简算法与 TARFD 算法进行对比分析,详细

结果见表 2~4. 其中 N_A 表示条件属性数目, A_{cc} 表示分类正确率.

表 2 给出了应用不同约简算法对数据集进行约简得到的属性子集,数字元代表了属性子集中与 UCI 数据集相对应属性的标号. 表 3 和 4 分别给出了在 JRip 和 C4.5 决策树条件下不同算法得到属性子集的分类精度比较情况.

表 2 不同约简算法下数据集属性子集的比较
Table 2 Comparison of reduction subsets under different reduction methods

数据集	Re	IG	FCBF	FDM	OR	TARFD
Glass	{3 4 8 7 2 1}	{4 3 6 7 8 2 1}	{3 4 8 6 7 2 1}	{1 2 3 4 , 5 6 7 8 9}	{4 7 2 6 1 3 8}	{1 3 7 4 2}
Ion	{24 3 8 5 14 , 7 16 34 29 9}	{5 6 33 29 3 , 21 34 8 13 7}	{5 7 6 28 33}	{3 4 5 9 , 16 31 32}	{8 22 34 27 , 32 6}	{3 1 5 7}
Isolet5	{395 394 71 , 103 461 426 135 , 462 105 72 396 , 73 104 106 460 , 453 455 454 459 , 456 457 458 70}	{462 461 73 , 460 458 107 , 457 456 74 , 459 106 105 , 75 72 453}	{462 461 460 , 459 458 73 457 , 106 456 455 453 , 454 107 105 74 , 395 103 70 71 , 72}	{8 47 325 , 387 403 410 , 477 509 608}	{584 547 395 , 394 549 43 516 , 454 413 107 397 , 458 414 515 391 , 412 456 455 , 411 453}	{302 325 403 , 410 477 8 73 , 69 387 461 166 , 467 269 19 , 46 47 509 608}
Iris	{3 4}	{3 4}	{3 4}	{1 2 3 4}	{3 4}	{3 4}
Vehicle	{8 18 7 12 , 9 3 10 11}	{12 7 8 , 11 9 3 6}	{12 7 , 8 11 9}	{1 2 5 6 10 , 12 15 16 18}	{9 12 8 7 11 3}	{3 8 10 17}
Water2	{7 13 31 12 30 , 18 20 6 17 10}	{13 7 37 31 , 12 6}	{7 13 37 , 6 31 12 20}	{1 13 17 , 29 30 36}	{37 13 6 7 12 , 31 20 14 22}	{1 15 10 7 , 11 27}
Wav	{11 10 6 12 9 , 15 14 17 13 5}	{15 7 8 , 16 6 14 17}	{15 7 8 , 16 14 6 17}	{1 3 8 10 , 11 15 20}	{15 8 13 7 6 18 , 14 11 17 9 16 10}	{20 4 9 13 16}

表 3 JRip 条件下约简算法之间的分类精度比较
Table 3 Comparison of classification accuracy under JRip classifier

数据集	Origin		Re		IG		FCBF		FDM		OR		TARFD	
	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$
Glass	9	68.69	6	65.89	7	68.69	7	68.69	9	68.69	7	68.69	5	64.01
Ion	34	89.74	10	88.89	10	90.31	5	89.74	7	90.88	6	85.19	4	90.03
Isolet5	617	66.71	23	45.29	15	29.31	20	38.42	9	31.30	20	53.43	18	46.18
Iris	4	94.00	2	95.33	2	95.33	2	95.33	4	94.00	2	95.33	2	95.33
Vehicle	18	68.56	8	64.89	7	65.60	5	49.88	9	64.54	6	57.33	4	61.82
Water2	38	82.73	10	83.49	6	83.69	7	82.73	6	82.92	9	83.88	6	81.57
Wav	21	76.10	10	75.60	7	68.10	7	68.10	7	73.00	12	74.70	5	68.20
average	105.86	78.08	9.86	74.20	7.71	71.58	7.57	70.41	7.29	72.19	8.86	74.08	6.29	72.45

表4 决策树 C4.5 条件下约简算法之间的分类精度比较

Table 4 Comparison of classification accuracy under C4.5 decision tree

数据集	Origin		Re		IG		FCBF		FDM		OR		TARFD	
	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$	N_A	$A_{cc}/\%$
Glass	9	66.82	6	66.82	7	69.63	7	69.63	9	66.82	7	69.63	5	71.96
Ion	34	91.45	10	93.16	10	92.30	5	88.32	7	90.60	6	84.90	4	91.17
Isolet5	617	78.06	23	51.12	15	34.96	20	47.53	9	39.90	20	61.06	18	52.21
Iris	4	96.00	2	96.00	2	96.00	2	96.00	4	96.00	2	96.00	2	96.00
Vehicle	18	72.46	8	70.45	7	66.43	5	56.97	9	71.16	6	66.55	4	67.61
Water2	38	90.00	10	81.57	6	83.11	7	83.30	6	82.34	9	83.87	6	83.30
Wav	21	74.30	10	72.60	7	69.60	7	69.60	7	72.20	12	73.70	5	71.20
average	105.86	81.30	9.86	75.96	7.71	73.15	7.57	73.05	7.29	74.15	8.86	76.53	6.29	76.21

一般情况下,未经约简的属性集理论上分类能力较强,准确率高,本文试验验证了这一点,因此主要比较 TARFD 和其他算法的约简结果。

对表 2~4 进行分析可得出:

1) 由表 2 中不同的约简子集可看出,与其他算法相比较,基于 TARFD 算法产生的约简结果平均属性子集最小。

2) 从表 3 的数据可看出,若采用 JRip 分类器,则应用 TARFD 算法产生的属性子集平均分类准确率较 Re 和 OR 算法略低,但此时属性个数明显较少;与 IG、FCBF 和 FDM 算法相比较,TARFD 算法用较少的平均属性获得了较高的分类准确率。

3) 由表 4 可知,采用 C4.5 决策树时,除 OR 算法外其他算法产生的属性子集分类准确率都低于 TARFD 算法。但是 OR 算法产生的平均属性数目明显多于 TARFD 算法,其差值达到 2.57,与除 Re 算法外的其他算法相比较,此差值最大;然而,即使 Re 算法得到的平均属性数目与 TARFD 算法相差 3.57,其得到的正确率仍然低于 TARFD 算法。

上述实验表明,TARFD 算法在得到最小属性约简子集的同时,分类性能并没有明显降低,相反,多数情况下较其他算法的分类正确率略有提高,造成这种现象的原因在于,TARFD 算法综合考虑属性与类别和属性之间的依赖度得出属性子集的重要性,在实现约简的同时最大程度上保证各属性相对分类的重要性,即该算法得到的属性子集可获得原属性集的绝大部分分类信息。因此,基于 TARFD 的属性约简算法不仅降低了属性维数,而且在一定程度上保证了分类器的正确率。

4 结论

1) 属性约简是模糊粗糙集理论的核心内容之一。通过综合考虑属性-类别和属性之间的模糊粗糙依赖关系可进一步确定属性的重要性和冗余性。

2) 所提方法克服了典型方法得到的约简结果中含有冗余属性的问题。

3) UCI 上数据集的仿真结果表明,本文方法能得到数据集的最小属性约简,同时在一定程度上保证了分类正确率。

参考文献:

- [1] PAWLAK Z. Rough set [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.
- [2] PAWLAK Z, SKOWRON A. Rudiments of rough sets [J]. Information Sciences, 2007, 177: 3-27.
- [3] CORNELIS C, JENSEN R, HURTADO G, et al. Attribute selection with fuzzy decision reducts [J]. Information Sciences, 2010, 180(2): 209-224.
- [4] 宋晶, 戚建淮. 基于云变换和特性关系下粗糙集的决策树构造 [J]. 西南交通大学学报, 2010, 45(2): 312-316.
SONG Jing, QI Jian-Huai. Decision tree construction based on cloud transform and rough set theory under characteristic relation [J]. Journal of Southwest Jiaotong University, 2010, 45(2): 312-316. (in Chinese)
- [5] 关欣, 衣晓, 何友. 一种新的粗糙集属性约简方法及其应用 [J]. 控制与决策, 2009, 24(3): 464-467.
GUAN Xin, YI Xiao, HE You. Knowledge reduction and its applications based on rough set [J]. Control and Decision, 2009, 24(3): 464-467. (in Chinese)

- [6] 张建军, 张静波. 一种新的基于粗糙集理论的决策表离散化方法 [J]. 西安电子科技大学学报, 2004, 31(3): 369-472.
ZHANG Jian-Jun, ZHANG Jing-Bo. A novel rough set theory-based decision table discretization algorithm [J]. Journal of Xidian University, 2004, 31(3): 369-472. (in Chinese)
- [7] BANERJEE M, PAL S K. Roughness of a fuzzy set [J]. Information and Computer Science, 1996, 93(3): 235-245.
- [8] DUBOIS D, PRADE H. Rough fuzzy sets and fuzzy rough sets [J]. Information and Computer Science, 1990, 17(2): 191-209.
- [9] 张贤达. 现代信号处理 [M]. 2 版. 北京: 清华大学出版社, 2002.
- [10] 王志勇, 郭创新, 曹一家. 基于模糊粗糙集和神经网络的短期负荷预测方法 [J]. 中国电机工程学报, 2005, 25(19): 7-11.
WANG Zhi-yong, GUO Chuang-xin, CAO Yi-jia. A method for short term load forecasting integrating fuzzy-rough set with artificial neural network [J]. Proceedings of the Chinese Society for Electrical Engineering, 2005, 25(19): 7-11. (in Chinese)
- [11] 徐菲菲, 苗夺谦, 魏莱, 等. 基于互信息的模糊粗糙集属性约简 [J]. 电子与信息学报, 2008, 30(6): 1372-1375.
XU Fei-fei, MIAO Duo-qian, WEI Lai, et al. Mutual information-based algorithm for fuzzy-rough attribute reduction [J]. Journal of Electronics & Information Technology, 2008, 30(6): 1372-1375. (in Chinese)
- [12] TSANG E C C, CHEN D G, YEUNG D S, et al. Attributes reduction using fuzzy rough sets [J]. IEEE Trans on Fuzzy Systems, 2008, 16(5): 1130-1141.
- [13] JENSEN R, SHEN Q. New approaches to fuzzy-rough feature selection [J]. IEEE Trans on Fuzzy Systems, 2009, 17(4): 824-838.
- [14] JENSEN R, SHEN Q. Fuzzy-rough sets assisted attribute selection [J]. IEEE Trans on Fuzzy Systems, 2007, 15(1): 73-89.
- [15] SHANG C J, BARNES D, SHEN Q. Facilitating efficient Mars terrain image classification with fuzzy-rough feature selection [J]. International Journal of Hybrid Intelligent Systems, 2011, 8(1): 3-13.
- [16] BHATT R B, GOPAL M. On fuzzy rough sets approach to feature selection [J]. Pattern Recognition Letters, 2005, 26(7): 965-975.
- [17] HU Q H, AN S, YU D R, et al. Soft fuzzy rough sets for robust feature evaluation and selection [J]. Information Sciences, 2010, 180(22): 4384-4400.
- [18] PISHARODY P K, VADAKKEPAT P, POH L A. Fuzzy-rough discriminative feature selection and classification algorithm, with application to microarray and image datasets [J]. Applied Soft Computing, 2011(11): 3429-3440.
- [19] BLAKE C, KEOGH E, MERZ C J, et al. UCI repository of machine learning databases [DB/OL]. [2006-09-28]. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- [20] COHEN W W. Fast effective rule induction [C] // 12th International Conference on Machine Learning, Tahoe City, July 9-12, 1995: 115-123.
- [21] QUINLAN J R. C4.5: programs for machine learning [M]. San Francisco: Morgan Kaufmann Publishers, 1993.
- [22] YU L, LIU H. Efficient Feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004(5): 1205-1224.

(责任编辑 杨开英)