一种基于一致性准则的属性约简算法

杨 明
(南京师范大学计算机科学与技术学院 南京 210097)
(江苏省信息安全保密技术工程研究中心 南京 210097)

摘 要 粗糙集方法提供了一种新的处理不精确、不完全与不相容知识的数学工具. 属性约简是粗糙集理论的重要研究内容之一,已有的大多数属性约简算法主要针对离散值属性的约简,面向连续值属性的约简报道较少. 为此,在引入新的对象一致性定义后,提出一种新的基于一致性准则的属性约简模型,该模型可针对离散或连续值属性进行有效的约简,是经典粗糙集属性约简模型的有效推广. 依据新模型,提出了一种基于一致性准则的属性约简算法,该算法可有效进行连续值属性的约简,且通过错分对象数的控制可有效增强属性约简的有效性. 理论分析和实验表明提出的算法是有效可行的.

关键词 粗糙集;属性约简;一致性准则
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2010.00231

A Novel Algorithm for Attribute Reduction Based on Consistency Criterion

YANG Ming
(*School of Computer Science and Technology, Nanjing Normal University, Nanjing* 210097)
(*Jiangsu Research Center of Information Security & Privacy Technology, Nanjing* 210097)

Abstract Rough set theory is a new mathematical tool to deal with imprecise, incomplete and inconsistent data. Attribute reduction is one of important parts researched in rough set theory. Many existing algorithms mainly aim at the reduction of discrete-valued attributes, very little work has been done for attribute reduction aiming to continuous-valued attributes. Therefore, in this paper, after introducing a new definition on consistency of objects, a novel model based on consistency criterion for attribute reduction is introduced. The newly designed model is very suitable for the decision table with discrete-valued or continuous-valued attributes, and an extension of the classical rough set model. Based on this model, a novel algorithm for attribute reduction based on consistency criterion is proposed. This algorithm can effectively obtain an attribute reduction for the decision table with continuous-valued attributes, and meanwhile the effectiveness of the attribute subset obtained by the new model can be enhanced by controlling the number of the misclassified or consistent objects. Theoretical analysis and experiments shows that the algorithm of this paper is efficient and feasible.

Keywords rough set; attributes reduction; consistency criterion

1 引 言

波兰数学家 Pawlak 在 20 世纪 80 年代初提出的 Rough Set(RS,粗糙集)是一种新的处理不精确、不完全与不相容知识的数学理论[1],近年来该理论在机器学习、数据挖掘及模式识别等多个领域得到了广泛的应用[2-4]. 在粗糙集理论中,属性约简是重

收稿日期:2008-11-30;最终修改稿收到日期:2009-09-23. 本课题得到国家自然科学基金(60873176)及江苏省自然科学基金(BK2008430)资助.杨 明,男,1964 年生,博士,教授,博士生导师,主要研究领域为数据挖掘、机器学习、粗集理论与应用. E-mail: m.yang@njnu.edu.cn.

要研究内容之一，也是知识获取的关键步骤. 现有的属性约简大体上可分为基于差别矩阵及其改进的属性约简算法[5-9]、基于正区域的属性约简算法[10-11]、基于启发式的属性约简算法[12-13] 3 种，但这些基于经典 Rough 集模型的算法都是针对离散值属性，不适合连续值属性的约简. 目前，在粗糙集理论框架下，进行连续值属性约简是当前粗糙集研究的重要内容之一，受到研究者的极大关注，取得了一定的进展[5,14-17]，但现有的模型主要侧重于理论模型，相应模型下的高效算法还报道不多，这是因为研究者重点放在经典粗糙集模型的改进，如文献[16]作者提出的覆盖粗糙集的拓扑方法.

从粒度计算角度来看，经典粗糙集除了需要连续属性值离散化外，采用等价关系对数据进行粒度划分不利于挖掘隐含在数据集的潜在知识，且人们认识自然也是从不同的粒度考虑的. 于是，研究者提出基于覆盖的方法，这些方法是对 Rough 集模型的有效扩展[14-17]. 这些扩展的属性约简模型为我们提供了理论依据，但相应的高效属性约简算法还报道不多. 为此，针对已提出的邻域覆盖模型[14,17]，文献[18]提出一种基于邻域的属性约简算法，以满足连续值属性的约简需要. 新模型诱导出的属性约简算法避免了连续值属性的离散化，增强了约简的有效性，为粗糙集模型下连续值属性的约简提高了一条新的途径. 但该算法须求解各对象的邻域，存在计算代价高的不足，且邻域参数选择缺乏理论上的合理解释.

为此，本文引入基于一致性准则的属性约简模型，该模型采用一致性对象保持策略来获取属性约简，不仅有效扩展了经典的粗糙集模型且可有效避免计算各对象的邻域，因而可有效提高属性约简的效率. 基于新模型，本文提出一种基于一致性准则的属性约简算法，该算法无须计算各对象的邻域，且依据统计机器学习的间隔理论[19-20]可对参数选择在理论上给出合理的解释. 实验结果表明本文算法是有效可行的.

2 粗糙集概念

粗糙集理论的要点是将分类与知识联系在一起，并用等价类关系形式化表示分类. 可理解为：知识是使用等价类 R 对离散空间 U 的划分，记为 $U/R=\{X_1,X_2,\cdots,X_n\}$，称为 X_i 为 U/R 的等价类. 为节省篇幅，仅介绍和属性约简及核有关的一些概念，关于粗糙集的其它一些概念可参见文献[2-3].

决策表 DT 是一个四元组 $\langle U,Q,V,f\rangle$，其中，U 是一组对象的非空有限集合，称为论域；设有 n 个对象，则 U 可表示为 $U=\{x_1,x_2,\cdots,x_n\}$，Q 是属性集合，$V=\bigcup_{a\in Q}V_a$，V_a 为属性 a 的值域集；f 是 $U\times Q\to V$ 的映射. 属性集合 Q 通常分为条件属性集 C 与决策属性集 D. 对 $B\subseteq Q$，无差别关系 $IND(B)$ 定义为 $\{(x,y)\in U^2 \mid \forall a\in B, f(x,a)=f(y,a)\}$，通过 $IND(B)$ 将 U 划分为若干个类 $E_i(1\leq i\leq |U/IND(B)|)$. 为便于叙述，设条件属性集合 C 中有 m 个属性 $C_1,C_2,C_3,\cdots,C_m$，其值域为有限离散集合，并用 $|\cdot|$ 表示集合的基. 不失一般性，假设仅有一个决策属性 D，其取值范围是 $1,2,\cdots,k$. 由 D 导出的等价类构成 U 的一个划分：$\{\psi_1,\psi_2,\cdots,\psi_k\}$，其中，$\psi_i=\{x\in U: f(x,D)=i\}$，$i=1,2,\cdots,k$.

定义 1[7]. 在决策表 DT 中，对 $P\subseteq C$，若两个不同的对象 x 和 y 在属性集 P 下具有相同的条件属性值而具有不同的分类，则称 x 和 y 是关于 P 不一致的，否则称 x 和 y 是关于 P 一致的.

定义 2. 设 $X\subseteq U$ 为论域的一个子集，$P\subseteq C$，X 的关于 P 的下近似为 $\underline{P}X=\{x\in U: [x]_P\subseteq X\}$；其中，$[x]_P$ 表示 U 中所有与 x 在关系 $IND(P)$ 下是等价的元素构成的集合.

定义 3. 设 $P\subseteq C$，对划分 $\{\psi_1,\psi_2,\cdots,\psi_k\}$ 的 P-近似精度为 $\gamma_P=\sum_{i=1}^{k}|\underline{P}\psi_i|/|U|$.

定义 4. 设 $P\subseteq C$，若 $\gamma_P=\gamma_C$，且不存在 $R\subset P$，使得 $\gamma_R=\gamma_C$，则称 P 为 C 的一个(相对于决策属性 D 的)属性约简. 称满足 $\gamma_P=\gamma_C$ 的条件属性子集 P 为候选属性约简. 所有 C 的属性约简的交称为 C 的核(简称核)，记为 $Core(C)$.

定义 5. 如果属性 $a\in C$ 满足 $\gamma_{C-\{a\}}<\gamma_C$，则称属性 a 为不可缺少的(indispensable)，否则，称属性为冗余的.

性质 1. 属性 $a\in Core(C)$ 当且仅当 a 是不可缺少的属性.

利用定义 4 和定义 5，研究者提出一些高效的属性约简算法(如文献[8,10-12]). 然而，这些基于经典 Rough 集模型的属性约简算法仅适用于离散值属性的约简. 因此，寻找可有效求解连续值属性的约简算法是本文的主要目标.

3 基于一致性准则的属性约简模型

为克服经典 Rough 集模型的不足，研究者对

Rough 集模型进行了扩展,使其适应连续值属性的约简,该方面的研究取得了一定的进展[14,17-18],但还有很多问题需要解决,如避免求解各对象的邻域、寻找邻域参数值在理论上的合理解释等. 为此,本节在引入对象的 ε-一致和ε-不一致概念后,提出基于一致性准则的属性约简模型并得到该模型下的若干性质.

定义 6. 设 DT 为一决策表,$P\subseteq C$,对两个对象 $x,y\in U$,$f(x,D)\neq f(y,D)$,若有 $d_P(x,y)>\varepsilon$ 或 $dissim_P(x,y)>\varepsilon$,则称 x,y 在 P 上是ε-一致的;否则,x,y 在 P 上是ε-不一致的,其中 $\varepsilon\geq 0$(ε 被称为一致性参数),$d_P(x,y)$ 或 $dissim_P(x,y)$ 表示两个对象 x 与 y 之间的距离或不相似度(相离度),如 $d_P(x,y)=\max\limits_{a\in P}|f(x,a)-f(y,a)|$.

定义 7. 设 DT 为一决策表,$P\subseteq C$,对 $x\in U$,若 $\forall y\in U$,$f(x,D)\neq f(y,D)$,有 $d_P(x,y)>\varepsilon$ 或 $dissim_P(x,y)>\varepsilon$,则称 x 在 P 上是ε-一致对象;否则,x 在 P 上是ε-不一致对象(非 ε-一致对象),其中,$\varepsilon\geq 0$,$d_P(x,y)$ 或 $dissim_P(x,y)$ 表示两个对象 x 与 y 之间的距离或不相似度(相离度). 为方便计,属性子集 $P(P\subseteq C)$ 上的所有 ε-一致对象集和所有 ε-不一致对象集分别简记为 $U(P,\varepsilon)$ 和 $IU(P,\varepsilon)$.

依据上述 ε-一致对象和ε-不一致对象的定义,得到决策表属性约简的新定义如下.

定义 8. 设 $DT=\langle U,C\cup D,V,f\rangle$ 为一决策表,设 $P\subseteq C,\varepsilon\geq 0$,若 $U(P,\varepsilon)=U(C,\varepsilon)$ 且 $U(O,\varepsilon)\subset U(P,\varepsilon)$($\forall O\subset P$),则称 P 是 C 的一个约简.

对定义 8,我们分离散值属性和连续值属性两种情况进行分析. 一方面,当决策表 DT 为离散值属性情况时,设 $\varepsilon=0$,若令 U 中两个对象在属性子集 P 上的距离 $d_P(x,y)$ 定义为

$$d_P(x,y)=\begin{cases}1, & \exists a\in P \text{ s.t. } f(x,a)\neq f(y,a)\\ 0, & \text{其它}\end{cases} \tag{1}$$

则令 $[x]_P=\{y|d_P(x,y)=0\}$ 可得与定义 2 一致的等价关系 $IND(P)$,从而可得 $|U(C,0)|/|U|=\gamma_C$,$|U(P,0)|/|U|=\gamma_P$. 可见,在特定约束下,定义 8 与经典属性约简的定义 4 是一致的. 因此,定义 8 是经典 Rough 集属性约简模型的推广.

另一方面,对连续值属性而言,设 $P\subseteq C,\varepsilon\geq 0$,令 $x(x\in U)$ 关于 P 的 ε-邻域为 $NN(x,P,\varepsilon)=\{y|d_P(x,y)\leq\varepsilon\}$. 对任意 $X\subseteq U$,定义 X 关于 P 的下近似为 $\underline{P}X=\{x\in U|NN(x,P,\varepsilon)\subseteq X\}$,有下列引理和定理成立.

引理 1. 若 $x_i(x_i\in U)$ 在 P 上是ε-一致对象,若 $f(x_i,D)=s$,则 $NN(x_i,P,\varepsilon)\subseteq\psi_s$.

证明. 反证法. 若存在 $y\in NN(x_i,P,\varepsilon)$,而 $y\notin\psi_s$,则存在 $j\neq s$,使 $y\in\psi_j$. 于是,有 $d_P(x_i,y)\leq\varepsilon$ 成立,这与 x_i 在 P 上是ε-一致对象矛盾. 故 $NN(x_i,P,\varepsilon)\subseteq\psi_s$. 证毕.

引理 2. 对给定的 $\varepsilon(\varepsilon\geq 0)$,设 $P\subseteq C$,若 $i\neq j(1\leq i,j\leq k)$,则 $\underline{P}\psi_i\cap\underline{P}\psi_j=\varnothing$.

证明. 反证法. 若 $\underline{P}\psi_i\cap\underline{P}\psi_j\neq\varnothing$,必存在 $x\in(\underline{P}\psi_i\cap\underline{P}\psi_j)$,则 $x\in(\psi_i\cap\psi_j)$,这与 $\psi_i\cap\psi_j=\varnothing$ 矛盾. 证毕.

定理 1. 对给定的 $\varepsilon(\varepsilon\geq 0)$,设 $P\subseteq C$,有 $U(P,\varepsilon)=\bigcup\limits_{i=1}^{k}\underline{P}\psi_i$ 成立.

证明. 因 $\underline{P}\psi_i=\{x\in U|NN(x,P,\varepsilon)\subseteq\psi_i\}$,由引理 1 和定义 7 可知 $U(P,\varepsilon)\subseteq\bigcup\limits_{i=1}^{k}\underline{P}\psi_i$ 成立. 反之,因 $\underline{P}\psi_i$ 中的每个对象都是 ε-一致对象,故 $\bigcup\limits_{i=1}^{k}\underline{P}\psi_i\subseteq U(P,\varepsilon)$ 成立. 证毕.

由定理 1 和引理 2 可知,有 $|U(P,\varepsilon)|=\sum\limits_{i=1}^{k}|\underline{P}\psi_i|$ 成立. 可见,由定义 8 给出的属性约简模型可诱导出基于邻域的属性约简模型但无须计算各对象的邻域. 也就是说,本文提出的基于一致性准则的属性约简模型仅关心不同类对象之间的可分性而无须关心同类对象之间的差别,这与统计机器学习的间隔理论是一致的. 依据间隔理论,若增大不同类之间的间隔,则间隔之间的误分对象数将增加;而依据本文基于一致性准则的属性约简模型,我们采用参数 ε 来有效控制误分对象数并使得间隔尽可能大. 为进一步分析参数 ε 与误分对象数之间的关系,我们给出下面的定理 2.

定理 2. 给定的 $\varepsilon_2\geq\varepsilon_1(\varepsilon_1\geq 0,\varepsilon_1\geq 0)$,对任意 $B\subseteq C$,有 $U(B,\varepsilon_2)\subseteq U(B,\varepsilon_1)$ 成立.

证明. 若 $x(x\in U)$ 在 B 上是 ε_2-一致对象,则对任意 $y\in U$,$f(x,D)\neq f(y,D)$,有 $d_B(x,y)>\varepsilon_2$,从而有 $d_B(x,y)>\varepsilon_1$,因此有 $U(B,\varepsilon_2)\subseteq U(B,\varepsilon_1)$ 成立. 证毕.

从定理 2 可以看出,增大参数 ε 意味着间隔的增大,同时意味着误分对象数可能增大. 寻找增大间隔且保持误分对象数尽可能少的属性子集是我们的一个主要目标. 此外,如何依据基于一致性准则的属性约简模型快速有效求解属性子集也是本文的另一个主要目标.

4 基于一致性准则的属性约简

4.1 基于一致性准则的属性约简算法

为快速有效地得到基于一致性准则的属性约简，需剖析属性子集不断扩展情况下一致对象集的变化(即其单调性).为此，引入下面的引理 3 和定理 3.

引理 3. 给定的 $\varepsilon(\varepsilon\geq 0)$，对任意 $A\subseteq B, B\subseteq C$，设对 $d_A(x,y)>\varepsilon$ 有 $d_B(x,y)>\varepsilon$ 成立，若 $x(x\in U)$是关于 A 的 ε-一致对象，则 $x(x\in U)$是关于 B 的 ε-一致对象.

事实上，常用的距离度量函数均满足引理 3 条件，如当属性子集 A 上的距离函数 $d_A(x,y)$定义为 p-范数($p=1,2,\infty$)时，引理 3 的条件成立，即若 $d_A(x,y)>\varepsilon$ 有 $d_B(x,y)>\varepsilon(A\subseteq B)$，简称距离度量 d_B满足单调性. 依据引理 3 可得下面的定理 3.

定理 3. 给定 $\varepsilon(\varepsilon\geq 0)$，设距离度量 d_B满足单调性，则对 $B1\subseteq B2\subseteq\cdots\subseteq Bn\subseteq C$，有 $U(B1,\varepsilon)\subseteq U(B2,\varepsilon)\subseteq\cdots\subseteq U(Bn,\varepsilon)\subseteq U(C,\varepsilon)$成立.

证明. 由引理 3 可知，对任意 $A\subseteq B, B\subseteq C$，由 $d_A(x,y)>\varepsilon$ 可得 $d_B(x,y)>\varepsilon$，即有 $U(A,\varepsilon)\subseteq U(B,\varepsilon)$成立. 故结论成立. 证毕.

依据定理 3，若给定的距离度量满足单调性，则通过逐步扩展重要属性即可得到一个有效的属性约简. 也就是说，对已得到的一致对象集 $U(B,\varepsilon)$，我们希望扩展这样的属性 $a\in(C-B)$使得 $U(B\cup\{a\},\varepsilon)$的一致对象数尽可能增多，同时希望寻找可快速求解 $U(B\cup\{a\},\varepsilon)$的策略. 为此目的，我们引入下面的定义 9、定理 4 和定理 5.

定义 9. 对给定的 $\varepsilon(\varepsilon\geq 0)$，$B\subseteq C$，定义区别矩阵 $\boldsymbol{M}_B=\{M_B(x_i,x_j)\}$为

$$M_B(x_i,x_j)=\begin{cases}1, & f(x_i,D)\neq f(x_j,D)\text{且}\\ & d_B(x_i,x_j)>\varepsilon,\\ 0, & \text{其它}\end{cases}\tag{2}$$

定理 4. 对给定的 $\varepsilon(\varepsilon\geq 0)$和属性子集 $B(B\subseteq C)$，其相应区别矩阵为 $\boldsymbol{M}_B$，若对 $x_i\in U, f(x_i,D)=s(1\leq s\leq k)$，有 $\sum_{j=1}^{|U|}M_B(x_i,x_j)=\sum_{j=1,j\neq s}^{k}|\phi_j|$成立，则 x_i在 B 上是 ε-一致对象；否则，x_i在 B 上是 ε-不一致对象.

证明. 由 $\sum_{j=1}^{|U|}M_B(x_i,x_j)=\sum_{j=1,j\neq s}^{k}|\phi_j|$知，$x_i$与任意不同类对象 $x_j\in U(f(x_j,D)\neq s)$在 B 上是 ε-一致的，因而由定义 7 可得 x_i在 B 上是 ε-一致对象. 否则，若 $\sum_{j=1}^{|U|}M_B(x_i,x_j)<\sum_{j=1,j\neq s}^{k}|\phi_j|$，则存在 $x_j\in U(f(x_j,D)\neq s)$使得 $M_B(x_i,x_j)=0$，即 $d_B(x_i,x_j)\leq\varepsilon$，从而 x_i与 x_j在 B 上是 ε-不一致的，进而 x_i在 B 上是 ε-不一致对象. 证毕.

为方便计，令 $I(M_B,x_i)=\sum_{j=1}^{|U|}M_B(x_i,x_j)$，$I(M_B)=\sum_{x_i\in U}I(M_B,x_i)$. 可以看出，$I(M_B)$越大，可区分的不同类对象数越多.

定理 5. 对给定的 $\varepsilon(\varepsilon\geq 0)$，设距离度量 d 满足单调性，若已知属性子集 $A(A\subseteq C)$和 $B(B\subseteq C)$的区别矩阵分别为 $\boldsymbol{M}_A$和 $\boldsymbol{M}_B$，则属性子集 $A\cup B$ 的区别矩阵 $\boldsymbol{M}_{A\cup B}$为

$$M_{A\cup B}(x_i,x_j)=\begin{cases}1, & M_A(x_i,x_j)=1\text{ 或 }M_B(x_i,x_j)=1,\\ 1, & f(x_i,D)\neq f(x_j,D)\text{且 }d_{A\cup B}(x_i,x_j)>\varepsilon,\\ 0, & \text{其它}\end{cases}\tag{3}$$

证明. 对任意两个不同类的对象 x_i和 x_j，若 $M_A(x_i,x_j)=1$ 或 $M_B(x_i,x_j)=1$，则因距离度量 d_B满足单调性，可得 $M_{A\cup B}(x_i,x_j)=1$；否则，对 $M_A(x_i,x_j)\neq 1$ 且 $M_B(x_i,x_j)\neq 1$，计算 $d_{A\cup B}(x_i,x_j)$，如果 $d_{A\cup B}(x_i,x_j)>\varepsilon$，那么 $M_{A\cup B}(x_i,x_j)=1$；否则，$M_{A\cup B}(x_i,x_j)=0$. 证毕.

依据定理 4，对给定的 $\varepsilon(\varepsilon\geq 0)$和属性子集 $B(B\subseteq C)$，可便捷快速求解 $U(B,\varepsilon)$. 而依据定理 5，可由两个属性子集上的区别矩阵 $\boldsymbol{M}_A$和 $\boldsymbol{M}_B$快速得到并集的区别矩阵 $\boldsymbol{M}_{A\cup B}$，从而由定理 4 可快速求解 $U(A\cup B,\varepsilon)$. 进一步，结合定理 3，可得如下属性重要性的评价准则：

$$Sig(a,B,\varepsilon)=|U(A\cup\{a\},\varepsilon)|-|U(B,\varepsilon)|\tag{4}$$

对式(4)，由定理 3 知 $Sig(a,B,\varepsilon)\geq 0$，若 $Sig(a,B,\varepsilon)=0$，则表明属性 a 相对于 B 来说重要性为 0，因而是冗余的. $Sig(a,B,\varepsilon)$的值越大表明属性 a 越重要. 因此，若采用前向搜索方法，则每次希望选择重要性尽可能大的属性，直到剩余的所有属性的重要性都为 0.

依据上述分析，我们可得一种基于一致性准则的属性约简算法，其主要思路为：(1) 对给定的决策表，初始化 $\varepsilon(\varepsilon\geq 0)$为一个合适的值，令约简集 $B=\varnothing$，选择一个满足单调性的距离度量函数 d 或不相似性度量 $dissim$；(2) 计算各属性 a 的区别矩阵；(3) 若 $(C-B)$中存在使 $Sig(a,B,\varepsilon)>0$ 最大的属性

a,则增加该属性 a 到 B;否则,若对 $a\in(C-B)$,$Sig(a,B,\varepsilon)$ 都为 0,则选择使得 $(I(\boldsymbol{M}_{B\cup\{a\}})-I(\boldsymbol{M}_B))$ 最大的属性 a 并增加到 B;(4)重复(3)直到 $U(B,\varepsilon)=U(C,\varepsilon)$ 为止.

依据上述一致性准则及相应的属性约简思路,基于一致性准则的属性约简算法的具体描述如下.

算法 1.　ARBCC(Attribute Reduction Based on Consistency Criterion).

输入:(1) $DT=\langle U,Q,V,f\rangle$;

(2)一致性参数 ε

输出:一个属性约简 R

主要步骤:

1. $\forall a\in C$:计算区别矩阵 $\boldsymbol{M}_a$;

2. $R=\varnothing$;

3. 计算 $U(C,\varepsilon)$;

4. 对 $\forall a\in(C-R)$,计算 $Sig(a,R,\varepsilon)$;

5. 求使 $Sig(a,R,\varepsilon)$ 最大的属性 $b=\arg\max\limits_{a\in(C-R)} Sig(a,R,\varepsilon)$;
//由定理 4 和式(4)

6. if $Sig(b,R,\varepsilon)>0$,
$R=R\cup\{b\}$;计算区别矩阵 $\boldsymbol{M}_R$;goto 4;
//由定理 5 可得

7. if $U(R,\varepsilon)\neq U(C,\varepsilon)$ then

7.1. 对 $\forall a\in(C-R)$,计算 $I(\boldsymbol{M}_{R\cup\{a\}})$;

7.2. 求使 $(I(\boldsymbol{M}_{R\cup\{a\}})-I(\boldsymbol{M}_R))$ 最大的属性
$b=\arg\max\limits_{a\in(C-R)}(I(\boldsymbol{M}_{R\cup\{a\}})-I(\boldsymbol{M}_R))$;

7.3. if $I(\boldsymbol{M}_{R\cup\{b\}})>I(\boldsymbol{M}_R)$ then
$R=R\cup\{b\}$;计算区别矩阵 $\boldsymbol{M}_R$;goto 4;
//由定理 5 可得

8. Return(R).

在 ARBCC 算法中,步 5 可由定理 4 和式(4)快速得到;步 6~7 中的 $\boldsymbol{M}_{R\cup\{b\}}$ 可由定理 5 快速求得,因而使得 ARBCC 算法可有效改进属性约简的效率.

由间隔理论可知,增大间隔可引起误分对象的增加,而增大一致性参数 ε 隐含着增大间隔,通过允许有一定数量的误分对象来合理增大间隔可有效增强分类器的推广性.我们知道一致对象数的减少意味着不一致对象的增加,而不一致对象的增加表明误分对象的增加,因而在 ARBCC 算法中通过控制不一致对象数可增强约简子集的有效性.因此,ARBCC 算法的有效性可从统计机器学习的间隔理论角度得到合理的解释,且参数 ε 的值可通过误分对象数来进行有效的设置;同时,也可通过 ε 来有效控制误分对象数.此外,为降低区别矩阵的空间代价可采用文献[6]的压缩存储策略;限于篇幅,该部分工作将另文讨论.为便于问题的讨论,本文的实验中采用的距离度量为 $d_B(x,y)=\max\limits_{a\in B}|f(x,a)-f(y,a)|$(即 ∞-范数).

4.2　与经典 Rough 集及邻域模型比较

与经典 Rough 集及邻域模型相比,本文提出的基于一致性准则的属性约简模型具有以下优点:

(1)经典 Rough 集属性约简模型仅适用于离散值属性的约简,即连续属性须先离散化,而基于一致性准则的属性约简模型既适用离散值属性的约简,也可直接用于连续值属性的约简.基于一致性准则的属性约简模型是经典 Rough 集属性约简模型的有效扩展.

(2)基于邻域的属性约简模型须计算各对象的邻域,而基于一致性准则的属性约简模型无须计算各对象的邻域,仅需计算不同类对象之间的相似性而不需要计算同类对象之间的相似性;同时,本文模型有效利用已得属性子集上的区别矩阵可快速求得两个属性子集的并集上的区别矩阵,因而使得由新模型诱导出的属性约简算法更加简洁快速.

从时间复杂度角度来看,对一个具有 N 个对象 k 类的决策表,其中各类对象数分别为 $N_1,N_2,\cdots,N_k$($\sum\limits_{i=1}^{k}N_i=N$),基于邻域的属性约简模型计算各对象邻域的时间复杂度至少为 $O(N^2)$;而基于一致性准则的属性约简模型计算不同类对象之间相似性的时间复杂度至多为 $O\left(\sum\limits_{1\leq i,j\leq k,i\neq j}N_iN_j\right)$,因而可有效提高计算效率.

(3)在基于一致性准则的属性约简模型中,参数 ε 的取值可从统计机器学习的间隔理论角度进行合理的解释,并可通过误分对象数加以合理控制,因而有利于增强分类器的推广性能.

实例 1.　假定数据集 dataset 是一个人工合成数据集,它是一个 2 维 2 类的数据集(第 1 维属性记为 a,第 2 维属性记为 b),其中第 1 类和第 2 类各有 60 个样本;第 1 类样本由高斯分布随机生成,其均向量为[3,1]、协方差阵为 $\begin{bmatrix}0.125 & 0\\ 0 & 0.125\end{bmatrix}$;第 2 类样本由均向量为[3,2.5]、协方差阵为 $\begin{bmatrix}0.125 & 0\\ 0 & 0.125\end{bmatrix}$ 的高斯分布随机生成,该数据集见图 1 所示.

为了有效解释参数 ε 的作用,本文给出一个简单的例子加以说明(见实例 1).采用经典 Rough 集属性约简算法均得到属性约简$\{a,b\}$.然而,若不考虑个别对象的误分,则属性 b 可将图 1 数据集中的

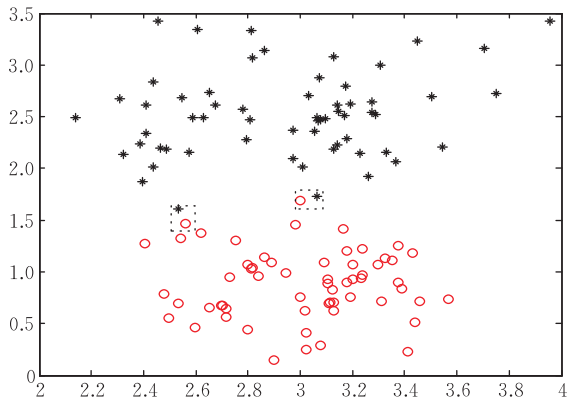


图 1 人工数据集

两类对象有效区分开来，如若允许图 1 中两个虚线框中的 4 个对象误分，则可得属性约简$\{b\}$；换句话说，若通过增大参数 ε 使得图 1 中两个虚线框中的 4 个对象为不一致对象，则有效增大两类之间的“间隔”，且可得到属性约简$\{b\}$.

当然，在实际应用中，如何有效选择参数 ε 是一个值得研究的问题. 本文的准则是希望找到“间隔”大且不一致对象数相对小情况下的参数 ε，该准则符合我们的直觉. 为了验证该准则的有效性，第 5 节将给出相关的实验结果.

5 实验结果

5.1 数据集描述

为进一步验证算法的性能，本文采用网上(http://www.ics.uci.edu)提供的 UCI 数据集，共有 11 个数据集，各数据集的描述见表 1.

表 1 实验中所采用的数据集描述

序号	数据集名称	对象数	类别数	条件属性数
1	Pima Indians Diabetes(Diabete)	768	2	8
2	Glass	214	6	9
3	Heart_disease(Hd)	270	2	13
4	Ionosphere	351	2	34
5	Iris	150	3	4
6	Vehicle	846	4	18
7	WBCD	683	2	9
8	WDBC	569	2	30
9	WPBC	194	2	32
10	Wine recognition data (Wine)	178	3	13
11	Waveform domain data (Wave)	5000	3	21

5.2 实验分析

为方便计，将文献[13]提出的属性约简算法简记为 Wang 算法，将文献[18]提出的基于邻域的属性约简算法简记为 Hu 算法. 我们对 ARBCC 算法、Wang 算法及 Hu 算法进行了性能比较. 对 ARBCC 算法和 Wang 算法，侧重比较他们的分类精度；而对 ARBCC 算法和 Hu 算法，主要比较他们的效率. 为便于讨论，参数 ε 的取值从集合$\{0.25,0.22,0.2,0.15,0.13,0.11,0.1,0.08,0.05,0.02,0\}$中选择，并采用“间隔”大且不一致对象数相对小的属性子集选择准则.

为测试 ARBCC 和 Wang 算法的性能，用表 1 中的前 10 个数据集来测试由属性子集诱导出的分类器精度. 在实验中，我们采用 10-fold 交叉验证的平均分类精度来评价属性子集的优劣，并分别用 3NN、C4.5、RBF 网络(简记为 RBFNN)及 KSVM 4 个不同的分类器来评价分类精度. 这里，KSVM 采用核化的 C-SVM，其核函数采用 RBF 核；分类器 3NN、C4.5、RBFNN 采用 Weka(Version 3-5)软件的缺省参数. ARBCC 和 Wang 的分类性能测试的实验结果如表 2 所示.

为测试 ARBCC 算法和 Hu 算法的效率，我们随机从 Wave 数据集中抽取 500，1000，1500，2000，3000 个对象构成一组实验数据，采用 Matlab7.0.4 实现这两种算法且比较参数 ε 取 0.15 和 0.25 两种情况下的算法执行效率，实验结果如图 2 所示.

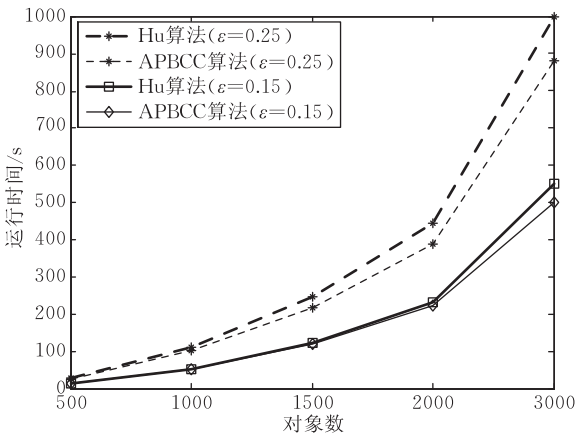


图 2 算法 ARBCC 和 Hu 的执行时间

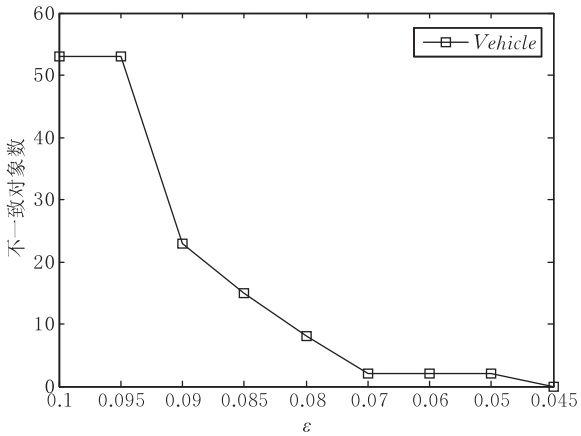
由图 2 可以看出，与 Hu 算法比，ARBCC 算法有效改进了属性约简的效率，这与 4.2 节的理论分析是一致的. 而图 3 和表 2 的实验结果则为参数 ε 的有效选择提供了一条新的途径，与 Hu 算法比，本文模型的参数值 ε 易于选择，具体说明如下：由图 3(a)可见，随着参数 ε 的逐渐减小，不一致对象数也逐渐减少，即“间隔”减小意味着误分对象减少，而“间隔”增大意味着误分对象逐步增多. 因此，我们希望在误分对象数较少的情况尽可能使“间隔”大. 而

由图 3(b)可见,当随着参数 ε 取值为 0.085 左右时,各分类器可取得比较满意的分类精度,而随着参数 ε 取值的不断减小,各分类器的精度是不断降低的,即对数据集 Vehicle, ε 应取[0.095,0.085]之间的某个值.该实验结果表明,在实际应用中,若当参数值 ε 较小时不一致对象数仍不为 0,则通过允许有一定数量不一致对象来增大参数 ε.

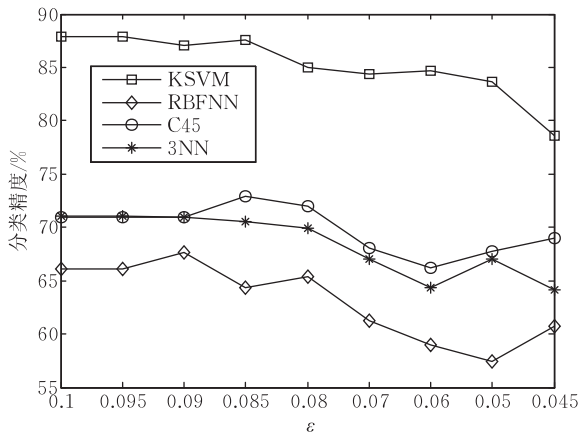
表 2　分类性能比较

数据集	算法	属性数	用 3NN 的精度	用 C4.5 的精度	用 RBFNN 的精度	用 KSVM 的精度
Diabete	Wang	-	73.8±4.8	72.4±5.1	73.9±3.9	80.7±4.7
	ARBCC(0.05)	6	72.1±4.0	**75.6±3.6**	73.7±4.8	80.5±3.8
Glass		-	69.0±8.0	69.7±4.8	62.2±8.1	79.4±3.7
	Wang	7	69.4±12.0	67.0±8.8	68.2±10.9	78.8±6.8
	ARBCC(0.02)	7	**72.0±8.0**	**72.8±6.0**	**71.9±10.3**	**82.3±7.4**
Hd		-	78.89±8.8	77.2±7.9	83.0±7.2	87.8±7.6
	Wang	11	76.7±11.4	79.3±7.4	**85.9±7.7**	**89.6±6.8**
	ARBCC(0.1)	8	**81.3±7.2**	**80.4±6.6**	81.5±7.4	87.8±9.7
Ionosphere		-	85.5±7.3	90.60±5.6	92.0±3.3	**98.3±3.6**
	Wang	3	84.9±7.1	85.2±4.8	79.2±8.0	91.5±5.4
	ARBCC(0.11)	10	**92.2±5.8**	**92.3±2.6**	**92.6±2.9**	97.4±2.4
Iris	Wang	-	94.7±6.5	94.0±5.5	96.0±6.8	98.7±4.3
	ARBCC(0.11)	2	**96.0±5.3**	**94.0±5.5**	**96.0±5.3**	**98.7±2.7**
Vehicle		-	**71.2±4.8**	72.2±3.7	64.8±6.1	**88.7±3.6**
	Wang	15	69.6±5.3	70.8±5.6	65.0±6.3	86.8±3.3
	ARBCC(0.08)	11	70.9±5.7	**72.2±3.3**	**65.4±6.7**	85.1±3.7
WBCD	Wang	-	96.2±2.6	95.60±3.4	96.1±2.5	98.2±2.0
	ARBCC(0.22)	7	**96.6±2.4**	**96.3±2.6**	**96.2±1.7**	**98.2±1.9**
WDBC		-	**97.1±2.4**	92.6±3.0	93.5±2.2	98.9±1.6
	Wang	8	93.0±2.3	91.6±2.6	93.8±2.5	97.5±2.1
	ARBCC(0.08)	11	96.5±2.0	**94.0±2.5**	**94.9±3.6**	**99.1±1.6**
WPBC		-	**71.5±6.7**	70.8±13.7	75.8±7.1	**84.0±6.5**
	Wang	2	67.6±6.0	**76.1±10**	75.5±10	82.4±7.3
	ARBCC(0.1)	8	71.0±12.5	76.1±10	**78.8±7.3**	82.4±8.0
Wine		-	95.9±2.7	90±6.5	97.1±3.9	99.4±1.8
	Wang	7	88.6±6.8	89.2±6.2	94.3±5.3	97.6±3.9
	ARBCC(0.15)	7	**97.6±2.9**	**92.9.1±5.2**	**97.6±2.9**	**99.4±1.8**

注:"-"表示条件属性集的基;黑色加粗表示较高识别率;ARBCC(0.15)表示当参数 ε 取 0.15 时得到的属性约简.这里的分类精度用 10-fold 交叉验证的平均分类精度和标准差来表示.



(a) 不一致对象数随参数ε的变化



(b) 分类精度随参数ε的变化

图　3

由表 2(表 2 中参数 ε 的值是在不一致对象数不超过 8 情况下得到的)可见,多数情况下,由 ARBCC 和 Wang 算法诱导出的分类器性能优于由全部属性集得到的分类器性能.进一步,对 Glass, Iris, Ionosphere, WBCD, WDBC 和 Wine 数据集而言,由 ARBCC 算法诱导出的分类器性能一致优于 Wang 算法诱导出的分类器性能;对 WPBC 数据集而言,由 ARBCC 算法诱导出的分类器性能与 Wang 算法

诱导出的分类器性能是可以比较的;对 Hd 和 Vehicle 而言,由 ARBCC 算法诱导出的分类器和 Wang 算法诱导出的分类器互有胜负;对 Diabete 而言,Wang 算法得到的属性子集是整个属性集,因而由 ARBCC 算法诱导出的分类器性能略优于 Wang 算法诱导出的分类器性能.可见,总体上 ARBCC 算法诱导出的分类器具有较优的分类性能.而某些情况下 ARBCC 算法诱导出的分类器性能略低,这可能是因为这里参数 ε 不是最优化得到的参数引起的.

综上分析,通过参数 ε 的选择,ARBCC 算法可有效降低约简属性集的规模,改进分类器的性能,因此是经典 Rough 集模型的拓展和改进.当然,在实际应用中,如何更加有效选择精度参数 ε,降低区别矩阵的存储代价将是我们的未来研究内容之一.

6 结 语

在引入新的一致性定义后,提出一种新的基于一致性准则的属性约简模型,该模型可针对离散或连续值属性进行有效的约简,是经典粗糙集属性约简模型的有效推广.依据新模型,提出了一种基于一致性准则的属性约简算法,该算法可有效进行连续值属性的约简,且通过错分对象数的控制可有效增强属性约简的有效性.理论分析和实验结果表明本文提出的算法是有效可行的.

参 考 文 献

[1] Pawlak Z. Rough sets. International Journal of Information and Computer Science, 1982, 11(5): 341-356

[2] Pawlak Z. Rough set approach to multi-attribute decision analysis. European Journal of Operational Research, 1994, 72(3): 443-459

[3] Liu Qing. Rough Sets and Rough Reasoning. Beijing: Science Press, 2001(in Chinese)
(刘清. Rough 集及 Rough 推理. 北京:科学出版社, 2001)

[4] Swiniarski R W, Skowron A. Rough set methods in feature selection and recognition. Pattern Recognition Letters, 2003, 24(6): 833-849

[5] Jensen R, Shen Q. Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1457-1471

[6] Yang Ming, Yang Ping. A novel condensing tree structure for Rough set feature selection. Neurocomputing, 2008, 71(4-6): 1092-1100

[7] Yang Ming. An incremental updating algorithm for attribute reduction based on improved discernibility matrix. Chinese Journal of Computers, 2007, 30(5): 815-822(in Chinese)
(杨明. 一种基于改进差别矩阵的属性约简增量式更新算法. 计算机学报, 2007, 30(5): 815-822)

[8] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks. Computational Intelligence, 1995, 11(2): 339-347

[9] Wang Jue, Wang Ju. Reduction algorithm based on discernibility matrix the ordered attributes method. Journal of Computer Science and Technology, 2001, 16(6): 489-504

[10] Liu Shao-Hui, Sheng Qiu-Jian, Wu Bin, Shi Zhong-Zhi, Hu Fei. Research on efficient algorithms for Rough set methods. Chinese Journal of Computers, 2003, 26(5): 524-529(in Chinese)
(刘少辉, 盛秋戬, 吴斌, 史忠植, 胡斐. Rough 集高效算法的研究. 计算机学报, 2003, 26(5): 524-529)

[11] Guan J W, Bell D A. Rough computational methods for information systems. Artificial Intelligences, 1998, 105(1-2): 77-103

[12] Miao Duo-Qian, Hu Gui-Rong. A heuristic algorithm for reduction of knowledge. Journal of Computer Research & Development, 1999, 36(6): 681-684(in Chinese)
(苗夺谦, 胡桂荣. 知识约简的一中启发式算法. 计算机研究与发展, 1999, 36(6): 681-684)

[13] Wang Guo-Yin et al. Theoretical study on attribute reduction of Rough set theory: Comparison of algebra and information views//Proceedings of the 3rd IEEE International Conference on Cognitive Informatics. Canada: IEEE Computer Society, 2004: 148-155

[14] Lin T Y. Computing on binary relation I: Data mining and neighborhood systems//Skoworn A, Polkowshi L eds. Proceedings of the Rough Sets in Knowledge Discovery. Physica-Verlag, 1998: 107-140

[15] Skowron A, Stepaniuk J. Information granules: Towards foundations of granular computing. International Journal of Intelligent Systems, 2001, 16(1): 57-85

[16] Zhu W. Topological approaches to covering Rough sets. Information Sciences, 2007, (177): 1499-1508

[17] Yao Y Y. Relational interpretations of neighborhood operators and Rough set approximation operators. Information Sciences, 1998, 111(1-4): 239-259

[18] Hu Qing-Hua, Yu Da-Ren, Xie Zong-Xia. Numerical attribute reduction based on neighborhood granulation and Rough approximation. Journal of Software, 2008, 19(3): 640-649(in Chinese)
(胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报, 2008, 19(3): 640-649)

[19] Crammer K, Gilad-Bachrach R, Navot A, Tishby N. Margin analysis of the lvq algorithm//Proceedings of the 17th Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2002: 1185-1192

[20] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

YANG Ming, born in 1964, Ph. D., professor, Ph. D. supervisor. His research interests include data mining and knowledge discovery, machine learning, rough sets theory and its applications.

Background

This work is supported by the National Natural Science Foundation of China under grant No. 60873176 and National Science of Jiangsu under grant No. BK2008430.

The main objective of this paper is to provide a novel framework for attribute reduction aiming to continuous-valued attributes. In this work, a new definition on consistency of objects is introduced, and a novel model is developed by the consistency criterion for attribute reduction. Based on this model, a novel algorithm for attribute reduction is proposed. From the viewpoint of margin theory, the author can effectively interrupt the effectiveness of the newly developed algorithm, since the effectiveness of the obtained attribute subset can be enhanced by controlling the number of the misclassified or consistent objects. Hence, the new model is an extension of the classical rough set model. Compared to the existing algorithms for the decision table with continuous-valued attributes or discrete-valued attributes, the new algorithm is efficient and feasible. How to improve the algorithm for further enhancing its efficiency is the author's future work.