基于优势-等价关系的几种约简及规则抽取方法

李 艳 孙娜欣 赵 津 王华超
(河北大学数学与计算机学院 河北省机器学习与计算智能重点实验室 保定 071002)

摘 要 考虑了条件属性引入优势关系而决策属性上引入等价关系的不协调目标信息系统。分析了这种基于优势-等价关系的相容约简、最大分布约简及正域约简三者之间的关系。此外,结合劣势关系抽取规则以提高规则的覆盖率,改进了基于优势关系的正域约简抽取规则(PDRIS)的方法。最后给出算例,并在UCI数据集上进行了大量的试验,以与PDRIS进行比较。

关键词 粗糙集,优势关系,等价关系,正域约简,规则抽取
中图法分类号 TP18 **文献标识码** A

Reductions Based on Dominance-Equivalence Relations and Rule Extraction Methods

LI Yan SUN Na-xin ZHAO Jin WANG Hua-chao
(Key Lab. in Machine Learning and Computational Intelligence, College of Mathematics and Computer Science, Hebei University, Baoding 071002, China)

Abstract Considered the inconsistent target information systems which respectively induce dominance relation on condition attribute set and equivalence relation on decision attribute set. Under this dominance-equivalence relation, the relationships were analyzed among three different reductions, compatible reduction, maximum distribution reduction and positive domain reduction. The rule extraction method based on positive domain reduction(PDRIS) was improved by introducing inferior relations together with dominance relations to extract rules with larger coverage. Finally, an example was given to illustrate our rule extraction method, and the experimental results on 13 UCI data were demonstrated to make comparisons between the proposed method and the PDRIS method.

Keywords Rough set, Dominance relation, Equivalence relation, Positive domain reduction, Rule extraction

1 引言

粗糙集理论[1-3]是近年发展起来的一种处理不精确、不确定及模糊信息的软计算工具。Pawlak提出的传统粗糙集理论是基于等价关系的,只能处理离散的符号值属性。它以完备的信息系统为研究对象,能够在保持分辨能力不变的前提下最大限度地进行知识约简。但实际问题中的属性值更多的是连续的或有偏序关系的,传统粗糙集必须先将属性值离散化才能进行处理,从而导致信息丢失,并且所得到的约简不能充分描述信息系统所涵盖的有序信息。许多学者对经典的粗糙集理论进行了改进[4-19],这些工作主要是考虑了属性值之间的关系,用优势关系替代等价关系,形成了基于优势关系的粗糙集理论。其中Greco等人[4-8]最早基于优势关系定义了优势类和上下联合的上、下近似;文献[9-18]针对条件和决策都是有序值属性的问题,定义了一系列不同约简,包括分配协调约简、近似协调约简、相容约简、下近似约简、上近似约简、分布约简和最大分布约简。这些约简在理论上推广了传统粗糙集理论,但很难在实际问题中解释它们的直观含义,也相应降低了所抽取规则的可解释性。并且,上述工作中的条件属性和决策属性都要求是有偏序关系的,而不少现实问题只有条件属性值是有序的,决策属性值则是没有优劣之分的。比如UCI数据库中的Iris,根据萼片长、宽以及花瓣的长和宽来判别花的种类(共有3种)。为方便叙述,以下把这种信息系统称为基于优势-等价关系的目标信息系统。如果这时也在决策属性上引入优势关系,那么会歪曲数据本身的实际意义,导致所抽取的规则对分类结果造成误导。针对这种情况,王国胤等人[19]提出了基于优势关系的不协调目标信息系统的正域约简。正域中的元素刻画的是:当某个对象的条件属性取值大于某个值时,它确定地属于某一类(而非优于某一类)。这种方法能够很好地反映上述基于优势-等价关系的信息系统所包含的信息。本文指出上面提到的基于优势关系的几种约简形式,可以通过将其决策优势类转化为决策等价类而成为基于优势-等价关系的约简,并讨论了优势-等价关系下的相容约简、目标信息系统的最大分布约简及正域约简3者之间的关系。在文献[19]给出的规则抽取方法的基础上,结合劣势关系对其进行改进,以提高规则的覆盖率,从而进一

到稿日期:2010-12-12 返修日期:2011-03-23 本文受国家自然科学基金(60903088),河北省自然科学基金(F2009000227,A2010000188,F2010000323),河北省第二批百名优秀人才支持计划,河北大学博士基金资助。

李 艳 博士,副教授,硕士生导师,主要研究方向为机器学习、Rough集理论、基于案例的推理、计算智能,E-mail:ly@hbu.cn;孙娜欣 硕士生,主要研究方向为Rough集理论、机器学习。

步提高对测试样例的识别率和正确识别率。

本文第 2 节给出优势关系、劣势关系等相关的基础知识；第 3 节给出优势-等价关系下的几种知识约简及其关系；在此基础上，第 4 节给出两种改进的规则抽取算法；第 5 节分析实验结果并与原有方法进行比较；最后给出结论。

2 基础知识

作为预备知识，本节介绍基于优势关系的不协调目标信息系统所涉及的相关概念，主要包括目标信息系统、等价关系、优势(劣势)关系及优势(劣势)类、集合的上下近似、正域及正域约简。

定义 1(目标信息系统)[9] 称一个五元组 $I=(U,A,F,D,G)$ 为一个目标信息系统，其中 (U,A,F) 是信息系统，A 称为条件属性集，D 称为目标属性集，F 是 U 与 A 的关系集，G 是 U 与 D 的关系集。

定义 2(等价及优(劣)势关系)[9] $I=(U,A,F,D,G)$ 为一个目标信息系统，对于 $B\subseteq A$，令

$$R_B=\{(x_i,x_j)\in U\times U: f_l(x_i)=f_l(x_j),\forall a_l\in B\}$$

$$R_B^{\preccurlyeq}=\{(x_i,x_j)\in U\times U: f_l(x_i)\leqslant f_l(x_j),\forall a_l\in B\}$$

$$R_B^{\succcurlyeq}=\{(x_i,x_j)\in U\times U: f_l(x_i)\geqslant f_l(x_j),\forall a_l\in B\}$$

R_B、$R_B^{\preccurlyeq}$、$R_B^{\succcurlyeq}$ 分别称为目标信息系统的等价关系、优势关系和劣势关系。

基于定义 2，记

$[x_i]_B=\{x_j\in U:(x_i,x_j)\in R_B\}=\{x_j\in U: f_l(x_i)=f_l(x_j),\forall a_l\in B\}$ 为 x_i 的等价类。

$[x_i]_B^{\preccurlyeq}=\{x_j\in U:(x_i,x_j)\in R_B^{\preccurlyeq}\}=\{x_j\in U: f_l(x_i)\leqslant f_l(x_j),\forall a_l\in B\}$ 为 x_i 的优势类。

$[x_i]_B^{\succcurlyeq}=\{x_j\in U:(x_i,x_j)\in R_B^{\succcurlyeq}\}=\{x_j\in U: f_l(x_i)\geqslant f_l(x_j),\forall a_l\in B\}$ 为 x_i 的劣势类。

定义 3(基于优势关系的集合的下近似及上近似)[9] 对于任意 $X\subseteq U$，定义 X 关于优势关系 $R_B^{\preccurlyeq}$ 的下近似和上近似分别为

$$\underline{R_B^{\preccurlyeq}}(X)=\{x_i\in U:[x_i]_B^{\preccurlyeq}\subseteq X\}$$

$$\overline{R_B^{\preccurlyeq}}(X)=\{x_i\in U:[x_i]_B^{\preccurlyeq}\cap X\neq\varnothing\}$$

本文以下所讨论的优(劣)势关系只体现在条件属性上，而决策属性还是基于等价关系，即只考虑基于优势-等价关系的目标信息系统。

下面介绍本文研究的目标信息系统的协调性。

定义 4(协调目标信息系统) 设 $I=(U,A,F,D,G)$ 为一个目标信息系统，若 $R_A^{\preccurlyeq}\subseteq R_D$，则称该基于优势-等价关系的目标信息系统是协调的，若 $R_A^{\preccurlyeq}\not\subseteq R_D$，则称该系统是不协调的。

定义 5(基于优势关系的正域及正域约简)[19] 设 $I=(U,A,F,D,G)$ 为一个目标信息系统，条件属性集相对于决策属性集的正域为 $POS_A(D)=\bigcup\limits_{X\in U/D}\underline{R_A^{\preccurlyeq}}(X)$。

对于 $B\subseteq A$，若 $POS_B(D)=POS_A(D)$，则 B 为正域协调集。若 B 为正域协调集，但 B 的任何真子集都不是正域协调集，则称 B 为正域约简。

同理，根据定义 2 可以类似给出基于劣势关系的正域及正域约简的概念，这里不再详细给出。

3 基于优势-等价关系的几种知识约简及其关系

文献[9-18]基于优势关系讨论了各种形式的约简，如分配协调约简、近似协调约简、相容约简、下近似约简、上近似约简、分布约简和最大分布约简，但是这些约简的提出是定义在条件属性和决策属性的值之间都是有序关系的目标信息系统上的。在现实生活中，有许多问题只是条件属性值是有序的，而决策属性值是没有优劣之分的。如前面提到的 Iris 数据集，其条件属性是连续的；而其决策属性有 3 类：Iris-setosa、Iris-versicolor 和 Iris-virginica，它们仅仅是鸢尾花的 3 个不同的种类，不能认为其中的一种比另外一种好。所以此时，要基于优势-等价关系来处理这类问题，那么上面提到的几种约简形式也就不再适用，而应将其决策优势类转化为等价类，从而得到基于优势-等价关系的分配协调约简、近似协调约简、相容约简、下近似约简、上近似约简、分布约简和最大分布约简。其中基于优势-等价关系的下近似约简也就是文献[19]中介绍的基于优势关系的正域约简。

下面首先把基于优势关系下提出的 3 种约简直接给出其对应于基于优势-等价关系的约简，讨论它们之间的关系并给出必要的证明。

3.1 基于优势-等价关系的相容约简与正域约简的关系

首先给出基于优势-等价关系的相容约简的定义。

定义 6(基于优势-等价关系的相容约简) 设 $I=(U,A,F,D,G)$ 为不协调目标信息系统，$B\subseteq A$，B 非空，若 $\overline{COM_B^{\preccurlyeq=}}=\overline{COM_A^{\preccurlyeq=}}$，则称 B 为优势-等价关系下的相容协调集。此外，若 B 的任意一个真子集都不是优势-等价关系下的相容协调集，则称 B 为优势-等价关系下的相容约简。

其中，$\overline{COM_A^{\preccurlyeq=}}=\{x_i\mid[x_i]_A^{\preccurlyeq}\subseteq[x_i]_D\}$

注：由定义可以看到，基于优势-等价关系的相容约简的本质是要保持那些满足 $[x_i]_A^{\preccurlyeq}\subseteq[x_i]_D$ 的对象的集合不变。

定理 1 基于优势-等价关系的相容协调集与正域协调集是等价的。

证明：根据集合下近似的定义和定义 5，可以看到

$$POS_A(D)=\bigcup_{X\in U/D}\underline{R_A^{\preccurlyeq}}(X)=\bigcup_{X\in U/D}\{x_i\mid[x_i]_A^{\preccurlyeq}\subseteq X\}\quad(1)$$

“⇒”设 B 为相容协调集，则有 $\overline{COM_B^{\preccurlyeq=}}=\overline{COM_A^{\preccurlyeq=}}$，对于 $\forall x\in\overline{COM_B^{\preccurlyeq=}}$，有 $[x]_B^{\preccurlyeq}\subseteq[x]_D$，$[x]_A^{\preccurlyeq}\subseteq[x]_D$，而 $[x]_D$ 为决策类，所以 $x\in POS_B(D)=POS_A(D)$，即 B 为正域协调集。

“⇐”设 B 为正域协调集，即有 $x\in POS_B(D)=POS_A(D)$ 和式(1)成立。若要证 B 为相容协调集，只需证明式(1)中的 X 为 x 自己的决策类。假设 $X\neq[x]_D$，则 $x\notin X$，这与 $[x]_A^{\preccurlyeq}\subseteq[x]_B^{\preccurlyeq}\subseteq X$ 矛盾，所以 B 为相容协调集。

综上，可知定理 1 成立。

由定理 1，易知有如下推论。

推论 1 基于优势-等价关系的相容约简与正域约简是等价的。

3.2 基于优势-等价关系的最大分布约简[17]与正域约简的关系

将文献[17]中的最大分布约简定义在优势-等价关系上，可以得到下面的定义。

定义 7(优势-等价关系下目标信息系统的最大分布约简)

设 $I=(U,A,F,D,G)$ 为基于优势-等价关系的目标信息系统，若对于 $B\subseteq A$，有 $\eta_A^{\preccurlyeq=}(I)=\eta_B^{\preccurlyeq=}(I)$，则称 B 是最大分布协调集，且 B 的任何真子集不是最大分布协调集，则称 B 为最大分布协调约简。

其中，

$$\eta_A^{\preccurlyeq=}(I)=\left\{x_i:x_i\in U \text{ 且} \frac{|[x_i]_A^{\preccurlyeq}\cap[x_i]_D|}{|[x_i]_A^{\preccurlyeq}|}=1\right\}$$

注：$\eta_A^{\preccurlyeq=}(I)=\{x_i:x_i\in U \text{ 且} [x_i]_A^{\preccurlyeq}\subseteq[x_i]_D\}$ 为 $\eta_A^{\preccurlyeq=}(I)$ 的恒等变形式，由此可以看出优势-等价关系下的目标信息系统的最大分布约简与相容约简是等价的。显然，可以得到下面的定理及推论。

定理 2 基于优势-等价关系的最大分布协调集与正域协调集等价。

推论 2 基于优势-等价关系的最大分布约简与正域约简是等价的。

此外，基于优势-等价关系的分配约简和近似约简仍然保持其在优势关系下的等价关系。

4 对基于正域约简抽取规则方法的改进

第 3 节定义和分析了基于优势-等价关系的目标信息系统中的几种约简，为提取相应的简化规则奠定了基础。本节首先简述正域约简方法（即 PDRIS 方法）的思想，随后提出结合劣势关系改进覆盖率和准确率的方法。

定义 8（属性子集的重要性）[19] 属性子集 B'（$B'\subset B\subseteq A$）的重要性表示为

$$SGF(B',B,D)=\frac{|POS_B(D)|-|POS_{B/B'}(D)|}{|U|}$$

当 $\forall b\in B-B'$，$SGF(\{b\},B-B',D)\neq 0$ 时，属性子集 $B-B'$ 是正域约简。从正域中抽取规则，规则的适用性可以用覆盖率来衡量。

定义 9（规则的覆盖率）[19] 规则的覆盖率定义为

$$\beta=|X|/|Y|$$

式中，$X=POS_B(D)$。

为了表达非正域的信息，去掉已获得规则的对象，对剩下的对象再次求取正域并约简，获取规则，直到所有对象都能成为正域中的元素为止。设置一个参数 α（称为规则的级数）来区别每次获取的规则，每循环一次，α 加 1。这样，所有对象最终均被不同级别的规则所覆盖。

样本识别时从 $\alpha=1$ 的规则开始。若在此 α 值的所有规则中能找到样本条件属性满足的规则，则不需再找 α 值更大的规则；若在同一 α 值的规则中，样本能够满足多条决策属性值不同的规则，则选择 β 最大的规则。

但是 PDRIS 方法抽取到的规则仅仅是优势规则，能够覆盖一部分样例，对于那些适合用劣势关系下的规则来描述的数据，是不能够给出决策的。比如学生评价问题，语文<=60 AND 数学<=60 THEN 该学生为“差生”。这些信息的忽略造成规则覆盖率不理想，从而拒识率较高，也会影响到正确识别率。对于正域约简的属性重要性算法及规则获取问题，为了降低其在测试集上的拒识率并提高其正确率，我们引入劣势关系，并基于劣势关系抽取类似上述学生评价的劣势规则。需要指出，虽然劣势关系只是将优势关系做简单修改就可以得到，但对于适合用这种规则描述的问题有着实用价值。引入劣势关系的方法有两种：基于优势＋劣势关系平行提取规则，以及基于优势劣势交叉提取规则。为了便于和 PDRIS 进行比较，依然基于属性重要度来寻找约简。

4.1 基于优势＋劣势的规则抽取

在训练集上，首先基于 PDRIS 获取规则；随后将该方法中的优势关系替换成劣势关系，重复获取规则。这样，我们就得到如下形式的带有级别 α 和覆盖率 β 的两组规则：

基于优势关系得到的规则的形式是

if $q_1\geqslant r_{q1}$ and $q_2\geqslant r_{q2}$ and $\cdots q_p\geqslant r_{qp}\rightarrow(d=d_t)(\alpha,\beta)$

基于劣势关系得到的规则的形式是

if $q_1\leqslant r_{q1}$ and $q_2\leqslant r_{q2}$ and $\cdots q_p\leqslant r_{qp}\rightarrow(d=d_t)(\alpha,\beta)$

注：$q_1,q_2,\cdots,q_p\in A$，$r_{q1},r_{q2},\cdots,r_{qp}$ 分别为它们的值，$d\in D$，d_t 为决策属性值。

在测试阶段，先用优势规则对整个数据集进行逐级匹配。随后把优势规则不能匹配（即拒识）的样例，用劣势规则进行逐级匹配，最终输出样例的类别。也就是将 PDRIS 方法不能识别的样例，用抽取的劣势规则再次进行识别。这样会一定程度降低拒识率，同时提高正确识别率。

4.2 基于优势与劣势交叉的规则抽取

在训练集上，逐级抽取规则，但是每一级规则都由两部分构成：基于优势关系得到的正域约简的规则和基于劣势关系得到的正域约简的规则。亦即在每一级上，先对整个数据集基于优势关系寻找正域约简、抽取规则，然后再对整个数据集基于劣势关系寻找正域约简、抽取规则（基于两种关系得到的规则的形式和 4.1 节所述相同）。仍然用 (α,β) 来度量规则的级数和覆盖率。级数越高，意味着匹配越复杂，而且用于获得这些规则的样例越少。在整个数据集上，这样的规则的可信度是很低的。基于此，我们期望能够降低规则的级数，而基于优势与劣势交叉的规则抽取方法，每一级上抽取到的规则不少于 PDRIS，而规则的级数不会高于 PDRIS。下面通过一个简单的例子来解释上面的方法。

例 1 表 1 是文献[19]中的目标信息系统，其中 $\{a,b,c\}$ 为条件属性集，$\{d\}$ 为决策属性。

表 1 一个基于优势-等价关系的目标信息系统

U	a	b	c	d
x_1	2	3	5	2
x_2	5	1.4	4	3
x_3	6	2.6	6	3
x_4	4	0.8	3	2
x_5	1	1	1	1
x_6	3	2	2	1

以下用基于优势与劣势交叉的方法来寻找约简，抽取规则。

(1)整个数据集上得到基于优势关系的正域为 $\{x_1,x_2,x_3\}$，正域约简为 $\{a,b\}$，抽取到的规则为

$R_1^{\preccurlyeq}:(a\geqslant 2)\wedge(b\geqslant 3)\rightarrow(d=2)|(\alpha=1,\beta=1/6)$

$R_2^{\preccurlyeq}:(a\geqslant 5)\wedge(b\geqslant 1.4)\rightarrow(d=3)|(\alpha=1,\beta=2/6)$

(2)基于劣势关系的正域为 $\{x_4,x_5,x_6\}$，用属性重要度的方法得到约简 $\{a,b\}$，抽取到的规则为

$R_4^{\succcurlyeq}:(a\leqslant 4)\wedge(b\leqslant 0.8)\rightarrow(d=2)|(\alpha=1,\beta=1/6)$

$R_6^{\succcurlyeq}:(a\leqslant 3)\wedge(b\leqslant 2)\rightarrow(d=1)|(\alpha=1,\beta=2/6)$

(3)将(1)、(2)中得到的正域中的元素从整个论域中去掉，剩下的信息系统包含零个对象，则获取规则的过程结束。从输出得到的规则，可以看到，使用基于优势与劣势交叉的方

法抽取到的规则均为一级规则(覆盖率为 100%)。而基于 PDRIS 抽取的规则为两条一级规则(覆盖率为 50%)和两条二级规则。

注:在(1)和(2)中得到的正域约简均为{a,b},这只是一个巧合。一般情况下,用优势关系和劣势关系得到的约简不一定是相同的。

(4)规则的匹配

在测试集上,样本识别时从 $\alpha=1$ 的规则开始。若在此 α 值的基于优势关系得到的规则中能找到样本条件属性满足的规则,则不需再找此 α 值的劣势规则,也不需要再找 α 值更大的规则;若在同一 α 值的规则中,样本能够满足多条决策属性值不同的规则,则选择 β 最大的规则。

5 实验结果与分析

选用 UCI 机器学习数据库中的 13 个数据集进行试验,这些数据集符合本文研究的对象的特征:条件属性都是连续属性,而且属性值之间有偏序关系;决策属性表示类别,都是离散值属性。本文提出的两种方法及文献[19]中 PDRIS 方法都在 MATLAB 上实现。随机取 50%作为训练集,其余样本作为测试集,对 5 次试验求平均值作为实验结果。结果如表 2 所列,其中斜体数字为每个数据集上最高正确识别率及最低拒识率。

表 2 实验结果对比

ID	数据集	N#	M#	优势+劣势		PDRIS		优势与劣势交叉	
				正确识别率	拒识率	正确识别率	拒识率	正确识别率	拒识率
1	sky	14	4	*0.7321*	*0.0410*	0.6125	0.0750	0.6875	0.2468
2	Ionosphere	351	34	0.7608	*0*	0.7080	0.1375	*0.7614*	0
3	shuttle	15	6	*0.8750*	*0*	0.8000	0.2000	*0.8750*	0
4	adult	20	7	*0.8727*	*0*	0.6545	0.3182	*0.8727*	0
5	Iris	150	4	*0.9107*	*0*	0.8880	0.0373	0.8100	0.0040
6	Diabetes	768	8	*0.7852*	*7.8125e-004*	0.7388	0.0674	0.6031	0.1187
7	Bupa	345	6	*0.6751*	*0*	0.6486	0.0474	0.6439	0.0780
8	Balance	625	4	*0.7773*	*0*	*0.7773*	0	0.6316	0.0230
9	Pima	768	8	*0.7664*	*7.8125e-004*	0.7044	0.0531	0.6041	0.1766
10	Auto-mpg	398	7	*0.6850*	*0.0870*	0.6100	0.2190	0.6105	0.1555
11	Machine	209	7	*0.7019*	*0*	0.6620	0.0852	0.6157	0.1167
12	Ecoli	336	7	*0.6653*	*0.0024*	0.5924	0.1112	0.5706	0.1988
13	Heart-statlog	270	11	*0.7148*	*0.0763*	0.5341	0.3126	0.5474	0.3496

注:在表 2 中均有 N#——样例个数,M#——条件属性个数。

正确识别率和拒识率的更加直观比较如图 3 所示(蓝色实线表示基于优势+劣势的方法的结果,绿色虚线表示 PDRIS 方法的结果,红色点线表示基于优势与劣势交叉方法得到的结果)。

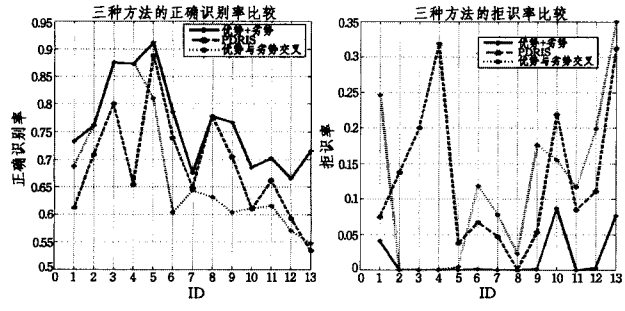

图 3 3 种方法的正确率和拒识率比较

图 3 左图表示基于 3 种方法的正确识别率,其中横坐标是数据集的 ID,纵坐标是正确识别率。可以看到,在这 13 个数据集上,实线总在虚线和点线之上。也就是说,通过优势+劣势的规则抽取方法得到的对测试集中样例的正确识别率要比 PDRIS 方法和优势与劣势交叉的规则抽取方法高,至少不低于后面两种方法。虚线和点线出现了交叉现象,也就是说在某些数据集上优势与劣势交叉的规则抽取方法得到的对测试集中样例的正确识别率要比 PDRIS 方法高;但是对某些数据集,前者却要低于后者。

图 3 右图表示基于 3 种方法的拒识率,其中横坐标是数据集的 ID,纵坐标是拒识率,节点表示该数据集基于某种方法得到的拒识率。可以看到,在这 13 个数据集上,实线总在虚线和点线之下。也就是说,通过优势+劣势的规则抽取方法得到的对测试集中样例的识别率要比 PDRIS 方法和优势与劣势交叉的规则抽取方法高,而且在大部分数据集上拒识率为零。虚线和点线仍然出现交叉现象,而在有些数据集上点线和横轴重合,也就是优势与劣势交叉的规则抽取方法得到的规则可以将测试集中的样例全部识别出来。

结束语 对于经典粗糙集不能很好处理的属性值有序关系的分类问题,用基于优势关系的粗糙集方法来处理更加合理。通过分析比较基于优势-等价关系下的相容约简、目标信息系统的最大分布约简以及正域约简,发现 3 者是等价的。对于仅基于优势关系得到约简抽取规则的方法,通过两种方法引入劣势关系来抽取规则(优势+劣势的规则抽取方法、优势与劣势交叉的规则抽取方法)。通过试验比较分析,我们发现优势+劣势的规则抽取方法得到的规则,对测试集中样例的识别率和正确识别率要比 PDRIS 方法和优势与劣势交叉的规则抽取方法高。但是比较优势与劣势交叉的规则抽取方法和 PDRIS,发现在 sky、Ionosphere、adult、shuttle、auto-mpg、heart-statlog 上,优势与劣势交叉的规则抽取方法的正确识别率要比 PDRIS 高,而在其他数据集上,却比较低。其原因可能是不同数据来源于不同领域的问题,属性的序信息有时适合用优势关系来体现,而有时适合劣势关系来表达。进一步研究包括考虑基于优势和等价混和关系的目标信息系统的约简及规则抽取。

参 考 文 献

[1] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. Boston: Kluwer Academic Publishers, 1991

[2] 苗夺谦,李道国. 粗糙集理论、算法与应用[M]. 北京:清华大学

出版社，2008
[3] Malcolm B. Reducts within the variable precision rough sets model: A further investigation [J]. European Journal of Operational Research, 2001, 134: 592-605
[4] Greco S, Matarazzo B, Slowingski R. Rough approximation of a preference relation by dominance relation [J]. European Journal of Operation Research, 1991, 117: 63-83
[5] Greco S, Matarazzo B, Slowinski R. A new rough set approach to evaluation of bankruptcy risk[M]. Zopounidis C. Ed. Operational Tool in the Management of Financial Risk. 1998: 121-136
[6] Greco S, Matarazzo B, Slowinski R. Dominance-based rough set approach to rating analysis[M]. Fuzzy Economic Review, 1999
[7] Greco S, Matarazzo B, Slowinski R. A new rough set approach to multicriteria and multiattribute classication[M]// Polkowski L, Skowron A. eds. Rough Sets and Current Trends in Computing, 1998: 60-67
[8] Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis [J]. European Journal of Operational Research, 2001, 129: 1-47
[9] 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
[10] 张文修，姚一豫，梁怡. 粗糙集与概念格[M]. 西安：西安交通大学出版社，2006
[11] 徐伟华，张文修. 基于优势关系下不协调目标信息系统的知识约简[J]. 计算机科学，2006，33(2)：182-184
[12] 袁修久，何华灿. 优势关系下的相容约简和下近似约简[J]. 西北工业大学学报，2006，24(5)：604-608
[13] 徐伟华，张晓燕，张文修. 优势关系下不协调目标信息系统的下近似约简[J]. 计算机工程与应用，2009，45(16)：66- 68
[14] 徐伟华，张晓燕，张文修. 优势关系下不协调目标信息系统的上近似约简[J]. 计算机工程，2009，35(18)：191-193
[15] 徐伟华，张文修. 基于优势关系下不协调目标信息系统的分布约简[J]. 模糊系统与数学，2007，21(4)：124-131
[16] 桂现才，彭宏. 优势关系下分布约简和最大分布约简问题研究[J]. 计算机工程与应用，2009，45(2)：150-153
[17] LI Yan, Sun Na-xin, Zhao Jin. Reductions in inconsistent decision systems based on dominance relations[C]// Proceedings of 2010 International Conference on Machine Learning and Cybernetics(ICMLC). 2010: 279-284
[18] LI Yan, Zhao Jin, Sun Na-xin, et al. Generalized Distribution Reduction in Inconsistent Decision Systems Based on Dominance Relations[C]//Proceedings of 2010 International Conference on Rough Sets and Knowledge Technology(RSKT). LNAI 6401, 2010: 151-158
[19] 陈娟，王国胤，胡军. 优势关系下不协调信息系统的正域约简[J]. 计算机科学，2008，35(3)：216-218

(上接第 205 页)

的比较。表 5 给出了超球算法和超椭球算法算法训练时间和分类时间的比较。

表 3 宏平均准确率、宏平均召回率和宏平均 F_1 值比较

算法	MAAP(%)	MAAR(%)	MAAF(%)
超球算法	78.38	77.92	77.52
超椭球算法	80.88	78.82	79.62

表 4 微平均准确率、微平均召回率和微平均 F_1 值比较

算法	兼类数	MIAP(%)	MIAR(%)	MIAF(%)
超球算法	1	71.34	73.91	72.14
	2	83.33	55.32	63.93
	3	100.00	50.00	65.00
超椭球算法	1	73.76	76.80	74.71
	2	85.19	59.26	71.67
	3	66.67	66.67	66.67

表 5 训练时间和测试时间比较

算法	训练时间(ms)	测试时间(ms)
超球算法	221	139
超椭球算法	276	101

从实验结果可以看出，超椭球方法的准确率和召回率都高于超球方法。这是因为样本在特征空间往往是带状的、凸的且各向异性的超椭球型分布，用超椭球包围的空间小于用超球包围的空间，从而提高了分类精度。超椭球方法较超球方法提高了分类速度，主要原因是每次分类时，超椭球方法的分类器中只涉及一个样本(待分类样本)，而超球方法的分类器涉及多个样本(所有支持向量)。超椭球方法的训练速度比超球方法略慢，这是因为构造超椭球时需要进行坐标变换，维数越高，计算量越大，同时还需优化缩放因子。

结束语 本文提出了一种基于超椭球的兼类文本类分类算法，描述了最小包围椭球的构造以及相应兼类文本分类算法，并与超球方法做了比较。在标准数据集 Reuters 21578 上的实验结果表明，该方法在分类精度和分类速度上都优于超球方法。进一步的研究工作是引入核函数理论，通过样本映射增加样本之间的紧密度，缩小超椭球包围空间，进一步提高分类精度。

参 考 文 献

[1] Yang Yi-ming. An Evaluation of Statistical Approaches to Text Categorization[J]. Journal of Information Retrieval, 1999(1): 69-90
[2] Mccallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification[C]// AAAI Workshop on Learning for Text Categorization. Madison, 1998: 509-516
[3] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. Beijing: Higher Education Press, 2001
[4] 林士敏，田凤占，陆玉昌. 贝叶斯学习、贝叶斯网络与数据采掘[J]. 计算机科学，2000，27(10)：69-72
[5] Takahashi F, Abe S. Decision-Tree-Based Multiclass Support Vector Machines[C]// International Conference on Neural Information Processing. Singapore, 2002: 1418-1422
[6] Platt J, Cristianini N, Shawe-Taylor J. Large Margin DAGs for Multiclass Classification[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 547-553
[7] Tax D, Duin R. Uniform Object Generation for Optimizing One-class Classifiers[J]. Journal of Machine Learning Research, 2001(2): 155-173
[8] 王晔，黄上腾. 基于支持向量机的文本兼类标注[J]. 计算机工程与应用，2006，42(2)：182-185
[9] 秦玉平，王秀坤，王春立. 实现兼类样本类增量学习的一种算法[J]. 控制与决策，2009，24(1)：137-140
[10] 高俊祥，杜海清，刘勇. 采用光照不变特征的椭球法运动阴影检测[J]. 北京邮电大学学报，2009，32(5)：109-113
[11] Shigeo A, Ruck T. A Fuzzy Classifier with Ellipsoidal Regions [J]. IEEE Transactions on Fuzzy Systems, 1997, 5(3): 358-368
[12] 刘勇，赵斌，夏绍玮. 模糊超椭球分类算法及其在无约束手写体数字识别中的应用[J]. 清华大学学报，2000，40(9)：120-124