# 扩展优势关系下的变精度粗糙集模型

李 艳 靳永飞 马红艳
(河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室 保定 071002)

**摘 要** 基于优势关系的变精度粗糙集模型将传统粗糙集中的等价关系扩展为优势关系,并结合变精度的思想来定义相关概念,从而可以处理具有偏好关系的信息并具有一定的容错能力。然而,传统优势关系的定义仍然过于严格,只有当一个对象 $x$ 的每个属性值都优于另一个对象 $y$ 时,该对象 $x$ 才优于 $y$。当属性个数较多时,这种优势关系的定义会导致对象的优势集偏小,影响到规则的提取和决策结果。为了解决这一问题,通过引入参数的方法扩展了传统优势关系的定义,并在此基础上进一步给出了扩展后的优势集和近似集的概念,建立了扩展优势关系下的变精度粗糙集模型,采用覆盖率和测试精度作为模型的评估指标。最后给出算例,并在 UCI 数据集上进行大量的实验将所提模型与传统优势关系下的变精度粗糙集模型进行比较。

**关键词** 优势关系,变精度粗糙集,扩展优势关系,近似集,决策规则
**中图法分类号** TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.046

## Variable Precision Rough Set Model Based on Extended Dominance Relations

LI Yan JIN Yong-fei MA Hong-yan
(Hebei Province Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Information Science, Hebei University, Baoding 071002, China)

**Abstract** The variable precision rough set (VPRS) model based on dominance relations extends equivalence relations in traditional rough sets to dominance relations, and combines with the idea of variable precision to define the relevant concepts. Therefore, it can deal with preference-ordered information with certain fault tolerance degree. However, the definition of traditional dominance relation is still too strict, in which object $x$ is superior to object $y$ only when all attribute values of $x$ are superior to that of $y$. This definition is difficult to be satisfied especially when the number of attributes is large. This will lead to smaller dominance, and even worse, it will affect the extraction of decision rules and then the process of decision making. To address this problem, the concept of dominance relation was extended by introducing a parameter and then the dominance set and approximation sets were correspondingly defined based on this extended dominance relation. The extended VPRS model was also developed, and the coverage rate and the test accuracy were used as evaluation criteria to for model. Finally, an illustrative example was given and the experimens on UCI data were conducted to compare the proposed extended model with the traditional VPRS model.

**Keywords** Dominance relation, Variable precision rough set, Extended dominance relation, Approximation sets, Decision rules

## 1 引言

随着科学技术的发展,特别是网络的普及和数据的广泛使用,如何从大量的数据中挖掘出有用的信息变得越来越重要。但是,实际应用中很多数据具有不确定性或不完整性,为了处理这类数据并从中找出有用的信息,波兰科学家 Z. Pawlak 提出了粗糙集理论(RS)[1-3]。RS 是一种优秀的处理不确定性数据的数学工具,现已经成功应用于模式识别、机器学习、数据挖掘、决策分析、医疗、经济等领域。RS 通过对关系数据库分类归纳形成概念和规则,利用等价关系形成精确的等价类来定义上、下近似集,处理数据的不精确性和知识的模糊性,亦即"包含"或"不包含",而没有某种程度上的"包含"或"属于"。

自粗糙集理论被提出以后,很多学者从不同角度对其进行了改进和扩展。为了克服下近似定义过于严格而上近似定义过于宽松的局限性,1993 年 Ziarko 提出了变精度粗糙集模型(VPRS)[4]。VPRS 是在 Pawlak 粗糙集模型中引入参数,放宽了上下近似的定义,即允许一定程度的错误分类率。这样既完善了近似空间的概念,又有利于用粗糙集理论从看似不相关的数据中发现潜在的相关数据[5],使得 RS 的应用领

到稿日期:2015-07-30 返修日期:2015-12-25 本文受国家自然科学基金(61170040,61473111),河北省自然科学基金(F2014201100,A2014201003),河北大学研究生创新资助项目(X2015059)资助。

李 艳(1976-),女,博士,教授,CCF 会员,主要研究方向为机器学习、Rough 集理论、计算智能,E-mail:ly@hbu.cn;靳永飞(1990-),女,硕士生,主要研究方向为 Rough 集理论、机器学习,E-mail:742215164@qq.com。

域变得更加广泛[6-8]。另一方面，基于等价关系的传统 RS 模型不能处理具有偏好关系的数据，例如对两个公司的破产风险进行评估时要依据各自负债率的“高低”；对学生整体的评价要依据各方面成绩的“优良差”。经典粗糙集理论只考虑属性值是否可区分，而不考虑它们的偏好关系，因此并不能在决策中很好地表达其原有的偏好信息。为了解决这个问题，Greco 等人在 1999 年首先提出了基于优势关系的粗糙集理论研究方法(DRSA)[9,10]。它用优势关系来取代等价关系，考虑对象之间的优劣信息，弥补了经典粗糙集理论在解决这类问题时的局限性，使其能够更好地解决实际问题[11-13]。

对经典粗糙集的扩展还包括 1987 年 Wong 等提出的概率粗糙集模型[14]。Yao 等通过将贝叶斯决策理论引入经典粗糙集，提出了决策粗糙集模型[15]。Dubois 等将被逼近的目标概念扩展为模糊集，提出了粗糙模糊集模型；将等价关系扩展为模糊相似关系，提出了模糊粗糙集模型[16]。A. Skowron 扩展等价关系为相容关系(相似关系)，提出了相容关系粗糙集模型[17]。

除此之外，Masahiro Inuiguchi 提出了变精度优势粗糙集模型[18-20]，即在经典粗糙集模型的基础上，将等价关系扩展为优势关系，并结合变精度的概念，对下近似进行了放宽，使其既能处理偏好信息，又有一定的容错能力。但此模型仍然采用传统的优势关系，要求一个对象的所有属性值均优于另一对象时这两个对象才满足优势关系的定义。这在实际问题中尤其是属性较多时很难得到满足，即满足优势关系的对象很少，造成优势集过小，从而进一步影响到近似集的大小和规则的提取。实际上，当一个对象 $x$ 在大多数属性上都优于另一对象 $y$ 时，通常就认为 $x$ 是优于 $y$ 的。例如，假设对甲、乙两所高校的学术声誉、学术资源、学术成果、学生情况、教师资源、物资资源的打分分别为：甲={90,89,80,87,85,82}；乙={85,83,79,86,82,83}。可以看到，在评估项中，甲只有物资资源稍劣于乙，按传统优势关系的概念，不能得出甲优于乙。但综合比较各项评估指标，应该可以得出甲高校是优于乙高校的。

为了放宽对传统优势关系的定义，本文引入参数 $\alpha$ ($0.5<\alpha\leqslant 1$)来定义扩展的优势关系和优势集，只要大部分的属性值优于另一对象的属性值，就说这个对象优于另一对象；并进一步提出基于扩展优势关系的变精度粗糙集模型，定义扩展优势关系下的变精度粗糙集的上下近似的概念。最后，基于近似集可以提取规则和进行决策。

本文第 2 节介绍了基本概念；第 3 节定义扩展优势关系，建立基于扩展优势关系的变精度粗糙集模型，给出算例来对比传统优势关系下的 VPRS 与扩展优势关系下 VPRS，结合覆盖率和测试精度对模型进行分析；第 4 节利用 12 组 UCI 中的数据验证扩展优势关系下 VPRS 的可行性与有效性；最后给出总结和讨论。

## 2 基本概念

作为预备知识，本节给出优势关系下的变精度粗糙集所涉及的几个基本概念，主要包括信息系统、优势关系及优势集、包含度、变精度粗糙集的上/下近似集。

**定义 1**(信息系统)[21] 四元组 $S=(U,A,V,f)$ 是一个信息系统，其中 $U$ 是对象的一个非空集合，称为论域；$A$ 为非空属性集合，$C$ 为条件属性集，$D$ 为决策属性集；$V$ 是属性值的集合；$f:U\times A\to V$ 是一个信息函数[22]，指在每个属性上对每个对象给出一个值，即 $\forall a\in A, x\in U, f(x,a)\in V$。

**定义 2**(划分)[23] 论域 $U$ 上的等价关系 $R$ 产生 $U$ 上的一个划分，记作 $U/R=\{[x_i]_R \mid x_i\in U\}$，其中 $[x_i]_R$ 为 $R$ 上的等价类。

**定义 3**(优势关系和优势集)[24] 设 $S=(U,A,V,f)$ 为一个信息系统，对于 $P\subseteq A$，令

$$D_P^+=\{(x_i,x_j)\in U\times U: f_l(x_i)\leqslant f_l(x_j), \forall a_l\in P\}$$

$D_P^+$ 称为连续值信息系统上的优势关系。根据定义 2，$D_P^+(x_i)=\{x_j\in U:(x_i,x_j)\in D_P^+\}=\{x_j\in U: f_l(x_i)\leqslant f_l(x_j), \forall a_l\in P\}$ 表示在属性集 $P$ 条件下优于对象 $x_i$ 的所有对象的集合，称为 $x_i$ 的优势集。

**定义 4**(包含度)[23] 令 $X$ 为全集 $U$ 中的一个非空子集，则 $c(x,X)=\dfrac{|D_P^+(x)\cap X|}{|D_P^+(x)|}$ 表示 $x$ 的优势集属于集合 $X$ 的包含度。其中，$|A|$ 表示集合 $A$ 中的元素个数，以下同理。

**定义 5**(VPRS 的上/下近似集)[17] 令 $X\subseteq U$，$0.5<\beta\leqslant 1$，分别称

$$\overline{R}_\beta^{\leqslant}(X)=\{x\in U:\frac{|D_P^+(x)\cap X|}{|D_P^+(x)|}\geqslant 1-\beta\}$$

$$\underline{R}_\beta^{\leqslant}(X)=\{x\in U:\frac{|D_P^+(x)\cap X|}{|D_P^+(x)|}\geqslant \beta\}$$

为 VPRS 的上、下近似集。

## 3 扩展优势关系下的变精度粗糙集模型

### 3.1 扩展优势关系和相关概念

本节首先定义扩展优势关系，对传统定义进行放宽，在此基础上进一步定义扩展后的优势集和上、下近似集，从而建立扩展优势关系下的 VPRS 模型。对扩展后的模型进行讨论和分析，得到的性质以命题的形式给出。

**定义 6**(扩展优势关系和优势集) 令 $S=(U,A,V,f)$ 是一个目标信息系统，$P\subseteq A$，$0.5<\alpha\leqslant 1$，$M=[\alpha\cdot m]$，$m$ 表示条件属性的个数，$[\cdot]$ 为向下取整符号。则定义

$$D_{P_\alpha}^+=\{(x_i,x_j)\in U\times U:\sum_{a_k\in P\wedge f(x_i,a_k)\leqslant f(x_j,a_k)}|\{a_k\}|\geqslant M\}$$

为信息系统 $S$ 上的扩展优势关系。在此基础上，定义 $D_{P_\alpha}^+(x_j)=\{x_j\in U:(x_i,x_j)\in D_{P_\alpha}^+\}=\{x_j\in U:\sum_{a_k\in P\wedge f(x_i,a_k)\leqslant f(x_j,a_k)}|\{a_k\}|\geqslant M\}$ 为 $x_i$ 在扩展 DRSA 下的优势集。

比较定义 6 和定义 3，传统的优势关系要求对象 $x_j$ 在 $P$ 中的每个属性 $a_l$ 上的属性值都大于或等于 $x_i$ 的属性值，$x_j$ 才属于 $x_i$ 的优势集，这对于大部分数据都难以满足；而扩展的优势关系放宽了要求，即只需 $x_j$ 在 $M=[\alpha\cdot m]$ 个属性上优于 $x_i$，就认为它属于 $x_i$ 的优势集。

**命题 1** 与传统优势关系相比，扩展优势关系下的优势集变大。

扩展后的优势关系放松了优势集的条件。若对象 $x_j$ 在 $s$ ($M\leqslant s<m$，$M=[\alpha\cdot m]$，$m$ 为所有属性的个数)个属性上优于 $x_i$，则在传统优势关系下它不属于 $x_i$ 的优势集；但在扩展的优势关系下满足优势集的定义。

**定义 7**(扩展优势关系下 VPRS 的上/下近似集) 令 $X\subseteq$

$U$，$0.5<\alpha\leqslant 1$，$0.5<\beta\leqslant 1$，记

$$\bar{R}_{\alpha\beta}^{\leqslant}(X)=\{x\in U:\frac{|D_{P_\alpha}^{+}(x)\cap X|}{|D_{P_\alpha}^{+}(x)|}\geqslant 1-\beta\}$$

$$\underline{R}_{\alpha\beta}^{\leqslant}(X)=\{x\in U:\frac{|D_{P_\alpha}^{+}(x)\cap X|}{|D_{P_\alpha}^{+}(x)|}\geqslant \beta\}$$

$c(x,X)=\dfrac{|D_{P_\alpha}^{+}(x)\cap X|}{|D_{P_\alpha}^{+}(x)|}$表示扩展后 $x$ 的优势集属于集合 $X$ 的包含度。分别称$\bar{R}_{\alpha\beta}^{\leqslant}(X)$与$\underline{R}_{\alpha\beta}^{\leqslant}(X)$为集合 $X$ 在扩展优势关系下变精度粗糙集的上、下近似集。

基于上述近似集的定义，可以进一步从中提取规则并进行决策。规则的一般形式为“如果对象 $x$ 优于对象 $y$，则 $x$ 以某包含度属于一类或多类”。这样，就建立了一个基于扩展优势关系的变精度粗糙集模型。为了与传统模型进行比较，引入覆盖率和测试精度两个评估指标，并对它们进行如下定义。

**定义 8**(覆盖率)　设 $x_i(i=1,2,\cdots,s)$为测试集中的对象，$x_k(k=1,2\cdots m)$为训练集中的对象。若 $\exists x_k$，使得 $x_i\in D_P^{+}(x_k)$，则称 $rate=n/s$ 为优势关系下的覆盖率，其中 $n=|x_i\in D_P^{+}(x_k)|$，即 $D_P^{+}(x_k)$中 $x_i$ 的数目。

**定义 9**(测试精度)　设测试集中对象 $x_i(i=1,2,\cdots,s)$的真实决策值为 $C(x_i)$；$X_j(j=1,2,\cdots,n)$为论域 $U$ 关于决策属性 $d$ 的划分；$x_k(k=1,2,\cdots,m)$为训练集中的对象，包含度 $\gamma_j=\dfrac{|D_P^{+}(x_k)\cap X_j|}{|D_P^{+}(x_k)|}$且 $1-\beta\leqslant\gamma_j\leqslant 1$。

(1)若 $x_i$ 匹配其中一条规则 $r_l$，具体形式有两种：

I. 如果 $x_i\in D_P^{+}(x_k)$，那么 $x_i\in X_j(\gamma_j)$；

II. 如果 $x_i\in D_P^{+}(x_k)$，那么 $x_i\in X_1(\gamma_1)$或 $x_i\in X_2(\gamma_2)$ 或$\cdots$ $x_i\in X_n(\gamma_n)(\sum_{j=1}^{n}(\gamma_j)=1$，且 $\gamma_j<\beta)$。

如果测试对象 $x_i$ 满足 I 且 $C(x_i)=X_j$，则 $x_i$ 能够准确地分类到 $X_j$，定义其准确率 $p_i=1$，否则 $p_i=0$；如果 $x_i$ 满足 II，且 $C(x_i)=X_j(j=1,2,\cdots,n)$，则认为 $x_i$ 能够准确分类的概率为 $\gamma_j$，定义其准确率 $p_i=\gamma_j$，否则 $p_i=0$。

(2)若 $x_i$ 不与任何一条规则 $r_l$ 匹配，则 $x_i$ 分类到每个类 $X_j$ 的概率相同，定义其准确率 $p_i=\dfrac{1}{n}$。称 $P=\dfrac{\sum_{i=1}^{s}p_i}{s}$为优势关系下变精度粗糙集的测试精度。

**命题 2**　与传统优势关系下变精度粗糙集方法相比，扩展优势关系下变精度粗糙集方法能够提高覆盖率，将更多的测试对象进行分类。

对于测试集中的对象 $x_k$，若在传统优势关系下其不属于任何训练对象的优势集 $D_P^{+}(x_i)$，即 $x_k$ 不能匹配任何一条规则，而利用扩展优势关系下变精度粗糙集方法所得的优势集明显扩大，这将会使 $x_k$ 可能属于某个 $D_{P_\alpha}^{+}(x_i)$而被规则所覆盖，从而能够有效提高规则的覆盖率。

### 3.2　与传统优势关系下变精度粗糙集模型的比较

本节给出一个简单的算例来说明将优势关系扩展后的变精度粗糙集模型在进行决策时的有效性。就覆盖率与测试精度两个方面对扩展优势关系下的 VPRS 与传统优势关系下的 VPRS 进行比较，同时讨论参数的优化选择问题。

选取的信息系统如表 1 所列，$x_1$ 到 $x_{15}$为训练集，$x_{16}$到 $x_{25}$为测试集。则论域 $U$ 关于决策属性 $d$ 的划分为：$U/d=\{X_1,X_2,X_3\}$，其中 $X_1=\{x_3,x_7,x_9,x_{13}\}$，$X_2=\{x_1,x_2,x_6,x_8,x_{10},x_{14},x_{15}\}$，$X_3=\{x_4,x_5,x_{11},x_{12}\}$。

表 1　一个信息系统

| U | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ | d |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 6.7 | 2.8 | 6.7 | 2 | 5.8 | 2.7 | 4.1 | 5.3 | 4.8 | 3.1 | 2.2 | 2 |
| $x_2$ | 5.1 | 2.5 | 3 | 1.1 | 5.7 | 4.4 | 1.5 | 0.8 | 4.4 | 3.1 | 1.8 | 2 |
| $x_3$ | 5.4 | 3.4 | 1.5 | 1.9 | 6.9 | 3.1 | 5.4 | 2.6 | 4.4 | 2 | 5.3 | 1 |
| $x_4$ | 5.1 | 3.4 | 1.5 | 0.2 | 6.1 | 2.6 | 5.6 | 3 | 5.4 | 0.5 | 6.3 | 3 |
| $x_5$ | 5.5 | 3.5 | 1.6 | 2 | 7.1 | 3.2 | 5.5 | 4.7 | 4.8 | 2 | 7.1 | 3 |
| $x_6$ | 6.3 | 2.5 | 3.2 | 1.3 | 5.8 | 2.8 | 4.2 | 6.2 | 4.8 | 4.2 | 2 | 2 |
| $x_7$ | 5.3 | 2.9 | 5.8 | 1.9 | 5.8 | 2.8 | 4.2 | 5.9 | 5.7 | 4.1 | 4.2 | 1 |
| $x_8$ | 6.3 | 3.2 | 3.1 | 1.2 | 6.9 | 4.5 | 1.8 | 4.3 | 5.3 | 4.3 | 4.3 | 2 |
| $x_9$ | 6.7 | 3 | 6.8 | 2.1 | 5.9 | 2.8 | 4.3 | 5.5 | 5 | 3.4 | 5.2 | 1 |
| $x_{10}$ | 6.7 | 3.1 | 6.9 | 2.2 | 5.7 | 2.6 | 3.9 | 6.5 | 6.9 | 2.3 | 1.8 | 2 |
| $x_{11}$ | 6.6 | 3 | 6.9 | 1.9 | 6.1 | 2.7 | 4.4 | 5.2 | 4.5 | 2.1 | 4.4 | 3 |
| $x_{12}$ | 6.7 | 3.6 | 6.9 | 2.2 | 7.1 | 3.3 | 5.6 | 6 | 5.7 | 4.2 | 7.1 | 3 |
| $x_{13}$ | 6.7 | 3.4 | 3.2 | 1.4 | 6.1 | 3.2 | 1.8 | 5.6 | 4.8 | 4.2 | 6 | 1 |
| $x_{14}$ | 6.2 | 2.7 | 5.9 | 1.8 | 5.7 | 2.8 | 4.1 | 6 | 5.8 | 6.8 | 6 | 2 |
| $x_{15}$ | 6.2 | 2.6 | 6.5 | 2.2 | 5.7 | 2.9 | 4.5 | 4 | 1.9 | 5.5 | 2.2 | 2 |
| $x_{16}$ | 5.3 | 3.1 | 6.2 | 2 | 5.8 | 2.9 | 4.2 | 6.3 | 5.7 | 4.3 | 4.2 | 1 |
| $x_{17}$ | 6.2 | 2.6 | 3 | 1.5 | 6.2 | 2.9 | 4.2 | 6.2 | 4.8 | 4.3 | 1.9 | 3 |
| $x_{18}$ | 6.8 | 2.9 | 6.5 | 1.9 | 5.8 | 2.8 | 4.2 | 5.3 | 4.8 | 3.2 | 2.3 | 1 |
| $x_{19}$ | 7.1 | 3 | 6.8 | 2.5 | 6.2 | 2.8 | 4.3 | 5.7 | 5.3 | 3.4 | 5.3 | 2 |
| $x_{20}$ | 5.5 | 3.4 | 1.6 | 2.1 | 6.2 | 3.2 | 5.6 | 2.8 | 4.5 | 1.9 | 5.4 | 1 |
| $x_{21}$ | 4.8 | 2.6 | 1.3 | 4.2 | 3.6 | 2.7 | 4.2 | 2.8 | 3.2 | 0.5 | 1.2 | 1 |
| $x_{22}$ | 5.5 | 3.2 | 1.6 | 2 | 6.8 | 3.2 | 5.5 | 2.8 | 4.8 | 2.3 | 5.4 | 2 |
| $x_{23}$ | 6.4 | 2.4 | 2.8 | 1.5 | 5.9 | 3.2 | 4.3 | 5.9 | 5 | 4.2 | 4.3 | 1 |
| $x_{24}$ | 4.3 | 2.3 | 3.2 | 0.2 | 4.2 | 1.9 | 1.3 | 2.8 | 3.9 | 2.5 | 3 | 2 |
| $x_{25}$ | 6.8 | 3.4 | 3.3 | 1.5 | 6.3 | 3.2 | 1.9 | 5.6 | 5 | 4.3 | 6 | 3 |

#### 3.2.1　传统优势关系下变精度粗糙集方法的覆盖率与测试精度

(1)计算优势类。

$D_P^{+}(x_1)=\{x_1,x_9,x_{12}\}$，$D_P^{+}(x_2)=\{x_2,x_8\}$，$D_P^{+}(x_3)=\{x_3,x_5,x_{12}\}$，$D_P^{+}(x_4)=\{x_4,x_{12}\}$，$D_P^{+}(x_5)=\{x_5,x_{12}\}$，$D_P^{+}(x_6)=\{x_6\}$，$D_P^{+}(x_7)=\{x_7,x_{12}\}$，$D_P^{+}(x_8)=\{x_8\}$，$D_P^{+}(x_9)=\{x_9,x_{12}\}$，$D_P^{+}(x_{10})=\{x_{10}\}$，$D_P^{+}(x_{11})=\{x_{11},x_{12}\}$，$D_P^{+}(x_{12})=\{x_{12}\}$，$D_P^{+}(x_{13})=\{x_{12},x_{13}\}$，$D_P^{+}(x_{14})=\{x_{14}\}$，$D_P^{+}(x_{15})=\{x_{15}\}$。

(2)取 $\beta=0.6$，计算 $X_i(i=1,2,3)$的上、下近似。

根据定义 5，则有：

$\underline{R}_{\beta}^{\leqslant}(X_1)=\varnothing$，$\bar{R}_{\beta}^{\leqslant}(X_1)=\{x_7,x_9,x_{13}\}$；

$\underline{R}_{\beta}^{\leqslant}(X_2)=\{x_2,x_6,x_8,x_{10},x_{14},x_{15}\}$，$\bar{R}_{\beta}^{\leqslant}(X_2)=\{x_2,x_6,x_8,x_{10},x_{14},x_{15}\}$；

$\underline{R}_{\beta}^{\leqslant}(X_3)=\{x_3,x_4,x_5,x_{11},x_{12}\}$，$\bar{R}_{\beta}^{\leqslant}(X_3)=\{x_3,x_4,x_5,x_7,x_9,x_{11},x_{12},x_{13}\}$。

(3)提取规则。

由(2)中得到的上、下近似，便可得到规则：

$r_1$：如果 $x\in D_P^{+}(x_2)$或 $x\in D_P^{+}(x_6)$或$x\in D_P^{+}(x_{10})$或$x\in D_P^{+}(x_{14})$或 $x\in D_P^{+}(x_{15})$，那么 $x\in X_2$；

$r_2$：如果 $x\in D_P^{+}(x_4)$或 $x\in D_P^{+}(x_5)$或 $x\in D_P^{+}(x_{11})$，那么 $x\in X_3$；

$r_3$：如果 $x\in D_P^{+}(x_3)$，那么 $x\in X_3(\frac{2}{3})$；

$r_4$：如果 $x\in D_P^{+}(x_7)$或 $x\in D_P^{+}(x_9)$或 $x\in D_P^{+}(x_{13})$，那么 $x\in X_1(\frac{1}{2})$或 $x\in X_3(\frac{1}{2})$。

(4)计算覆盖率与测试精度。

观察测试集$\{x_{16},x_{17},x_{18},x_{19},x_{20},x_{21},x_{22},x_{23},x_{24},x_{25}\}$中的对象，有：$x_{16}\in D_P^+(x_7)$，则$x_{16}\in X_1(\frac{1}{2})$或$x_{16}\in X_3(\frac{1}{2})$；$x_{19}\in D_P^+(x_9)$，则$x_{19}\in X_1(\frac{1}{2})$或$x_{19}\in X_3(\frac{1}{2})$；$x_{25}\in D_P^+(x_{13})$，则$x_{25}\in X_1(\frac{1}{2})$或$x_{25}\in X_3(\frac{1}{2})$。

即对象$x_{16},x_{19},x_{25}$可能被正确地分类，而对象$x_{17},x_{18},x_{20},x_{21},x_{22},x_{23},x_{24}$不能够分类。从而覆盖率$rate=\frac{3}{10}$，测试精度$P=\frac{\sum_{i=1}^{s}p_i}{s}=\frac{\sum_{i=16}^{25}p_i}{10}=\frac{\frac{1}{2}\times2+0+\frac{1}{3}\times7}{10}=\frac{1}{3}$。

### 3.2.2 扩展优势关系下变精度粗糙集方法的覆盖率与测试精度

(1)计算优势集。

取$\alpha=0.8$，$M=[\alpha\cdot m]=[0.8*11]=[8.8]=8$。

$D_{P_\alpha}^+(x_1)=\{x_1,x_7,x_9,x_{12},x_{13}\}$，$D_{P_\alpha}^+(x_2)=\{x_1,x_2,x_3,x_5,x_6,x_7,x_8,x_9,x_{10},x_{11},x_{12},x_{13},x_{14},x_{15}\}$，$D_{P_\alpha}^+(x_3)=\{x_3,x_5,x_{12},x_{13}\}$，$D_{P_\alpha}^+(x_4)=\{x_4,x_5,x_{12},x_{13}\}$，$D_{P_\alpha}^+(x_5)=\{x_5,x_{12}\}$，$D_{P_\alpha}^+(x_6)=\{x_6,x_7,x_9,x_{12},x_{13}\}$，$D_{P_\alpha}^+(x_7)=\{x_7,x_9,x_{12}\}$，$D_{P_\alpha}^+(x_8)=\{x_8,x_{12}\}$，$D_{P_\alpha}^+(x_9)=\{x_9,x_{12}\}$，$D_{P_\alpha}^+(x_{10})=\{x_{10},x_{12}\}$，$D_{P_\alpha}^+(x_{11})=\{x_9,x_{11},x_{12},x_{13}\}$，$D_{P_\alpha}^+(x_{12})=\{x_{12}\}$，$D_{P_\alpha}^+(x_{13})=\{x_{12},x_{13}\}$，$D_{P_\alpha}^+(x_{14})=\{x_{12},x_{14}\}$，$D_{P_\alpha}^+(x_{15})=\{x_{12},x_{15}\}$。

(2)计算$X_i(i=1,2,3)$的上下近似。

根据定义7，有

$\underline{R}_{\alpha\beta}^{\leqslant}(X_1)=\{x_1,x_6,x_7\}$，$\overline{R}_{\alpha\beta}^{\leqslant}(X_1)=\{x_1,x_3,x_6,x_7,x_9,x_{11},x_{13}\}$；$\underline{R}_{\alpha\beta}^{\leqslant}(X_2)=\varnothing$，$\overline{R}_{\alpha\beta}^{\leqslant}(X_2)=\{x_2,x_8,x_{10},x_{14},x_{15}\}$；$\underline{R}_{\alpha\beta}^{\leqslant}(X_3)=\{x_4,x_5,x_{12}\}$，$\overline{R}_{\alpha\beta}^{\leqslant}(X_3)=\{x_3,x_4,x_5,x_8,x_9,x_{10},x_{11},x_{12},x_{13},x_{14},x_{15}\}$。

(3)提取规则。

由(2)中得到的上、下近似，便可得到规则：

$r_1$：如果$x\in D_{P_\alpha}^+(x_5)$，那么$x\in X_3$；

$r_2$：如果$x\in D_{P_\alpha}^+(x_4)$，那么$x\in X_3(\frac{3}{4})$；

$r_3$：如果$x\in D_{P_\alpha}^+(x_7)$，那么$x\in X_1(\frac{2}{3})$；

$r_4$：如果$x\in D_{P_\alpha}^+(x_1)$或$x\in D_{P_\alpha}^+(x_6)$，那么$x\in X_1(\frac{3}{5})$；

$r_5$：如果$x\in D_{P_\alpha}^+(x_2)$，那么$x\in X_2(\frac{1}{2})$；

$r_6$：如果$x\in D_{P_\alpha}^+(x_3)$或$x\in D_{P_\alpha}^+(x_{11})$，那么$x\in X_1(\frac{1}{2})$或$x\in X_3(\frac{1}{2})$；

$r_7$：如果$x\in D_{P_\alpha}^+(x_8)$或$x\in D_{P_\alpha}^+(x_{10})$或$x\in D_{P_\alpha}^+(x_{14})$或$x\in D_{P_\alpha}^+(x_{15})$，那么$x\in X_2(\frac{1}{2})$或$x\in X_3(\frac{1}{2})$。

(4)计算覆盖率与测试精度。

观察测试集中的对象，根据扩展优势集的定义，有：$\{x_{20},x_{25}\}\subseteq D_{P_\alpha}^+(x_4)$，则$x_{20}\in X_3(\frac{3}{4})$，$x_{25}\in X_3(\frac{3}{4})$；$\{x_{16},x_{19}\}\subseteq D_{P_\alpha}^+(x_7)$，则$x_{16}\in X_1(\frac{2}{3})$，$x_{19}\in X_1(\frac{2}{3})$；$x_{18}\in D_{P_\alpha}^+(x_1)$，则$x_{18}\in X_1(\frac{3}{5})$；$\{x_{17},x_{23}\}\subseteq D_{P_\alpha}^+(x_6)$，则$x_{17}\in X_1(\frac{3}{5})$，$x_{23}\in X_1(\frac{3}{5})$；$x_{22}\in D_{P_\alpha}^+(x_2)$，则$x_{22}\in X_2(\frac{1}{2})$。

即对象$x_{16},x_{17},x_{18},x_{19},x_{20},x_{22},x_{23},x_{25}$可能被正确地分类，对象$x_{21},x_{24}$不能够分类，从而覆盖率$rate=\frac{8}{10}=\frac{4}{5}$，测试精度$P=\frac{\sum_{i=1}^{s}p_i}{s}=\frac{\sum_{i=16}^{25}p_i}{10}=\frac{1\times5+0\times3+\frac{1}{3}\times2}{10}=\frac{17}{30}$。

通过上述算例比较两种模型的优势集与覆盖率，很容易看到扩展后的优势集明显变大，覆盖率也大大提高，测试集中可能被正确分类的对象增加，验证了命题1与命题2。

$0.5<\alpha\leqslant1$，$\alpha$的变化可能会影响到覆盖率与测试精度。下面依次考虑当$\alpha=0.55,0.6,0.65,\cdots,1$时的情况，方法与$\alpha=0.8$时的相同，便得到$\beta=0.6$不变时，在不同的$\alpha$值下覆盖率与测试精度的对比结果，如表2所列。

根据表2，在不同的$\alpha$值下，综合考虑覆盖率与测试精度可以得到：当$\alpha=0.75$或$\alpha=0.8$时最为合适，即此时能够匹配更多的规则。

表2 $\beta=0.6$不变而$\alpha$变化时的覆盖率与测试精度的对比

| α | 0.55 | 0.6 | 0.65 | 0.7 | 0.75 | 0.8 | 0.85 | 0.9 | 0.95 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 覆盖率 | 0.8 | 0.8 | 0.8 | 0.8 | **0.8** | **0.8** | 0.7 | 0.7 | 0.5 | 0.3 |
| 测试精度 | 0.4667 | 0.4667 | 0.1667 | 0.1667 | **0.5667** | **0.5667** | 0.1 | 0.1 | 0.2667 | 0.3333 |

**命题3** 当$\alpha=1$时，扩展优势关系退化为传统优势关系，即传统优势关系下变精度粗糙集模型为扩展优势关系下变精度粗糙集模型的特殊情况。

另外，$0.5<\beta\leqslant1$，当取定一个最优$\alpha(\alpha=0.8)$时，$\beta$变化时可能也会影响到覆盖率与测试精度。下面依次考虑$\beta=0.55,0.6,0.65,\cdots,1$时的覆盖率与测试精度，方法与$\beta=0.6$时的相同，便得到$\alpha=0.8$不变时，在不同的$\beta$值下覆盖率与测试精度的比较结果如表3所列。

表3 $\alpha=0.8$不变而$\beta$变化时的覆盖率与测试精度的对比

| β | 0.55 | 0.6 | 0.65 | 0.7 | 0.75 | 0.8 | 0.85 | 0.9 | 0.95 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 覆盖率 | **0.8** | **0.8** | **0.8** | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| 测试精度 | **0.5667** | **0.5667** | **0.5667** | 0.5333 | 0.4583 | 0.2083 | 0.2083 | 0.2083 | 0.2083 | 0 |

综合考虑覆盖率与测试精度，可以得到当$\beta=0.55$，$\beta=0.6$或$\beta=0.65$时结果最优，即在覆盖率相同的情况下所能正确分类的测试对象最多。

注1：当取$\beta=1$时，模型退化为扩展优势关系下传统粗

糙集模型，即扩展优势关系下传统粗糙集模型为扩展优势关系下变精度粗糙集模型的特殊情况。

注 2：与扩展优势关系下传统粗糙集方法相比，采用扩展优势关系下变精度粗糙集方法时测试精度提高，所能确定的测试集中的对象增加。

## 4 实验分析

为了验证所提出的扩展优势关系下的变精度粗糙集方法的有效性，从 UCI 数据集中选取了 12 个数据集，这些数据集符合本文研究对象的特征：条件属性是具有偏好关系的，而决策属性是符号值属性。因此，在条件属性上定义优势关系，而在决策属性上采用等价关系对论域进行划分。对于每一个数据集，随机选取 60%为训练集，其余的为测试集。将对传统优势关系下的变精度模型与扩展优势关系下的变精度粗糙集模型的覆盖率与测试精度进行比较。所有的算法是在 Windows7，Intel（R） Core（TM）i3-4150 CPU @ 3.50 GHz 3.50GHz 4.00GB 环境下通过 Matlab（7.0）实现的。由于优势集的计算限制在完备的信息系统中，因此删去了数据集中没有属性值的对象。

对 UCI 上的数据集首先利用与第 3 节相同的方法，固定 $\beta=0.6$，综合考虑覆盖率与测试精度取得最优的 $\alpha$；然后取所得到的最优 $\alpha$ 作为扩展优势关系下变精度粗糙集模型的参数值并将其与传统优势关系下变精度粗糙集模型进行比较。表 4 列出了在 12 组数据上的实验结果。可以看出，在绝大多数情况下，扩展后的模型无论在覆盖率还是测试精度上均明显优于传统模型。如传统模型的覆盖率和测试精度在 12 组数据上的平均值分别为 0.2395 和 0.4524；而扩展后的模型的这两项指标平均值分别为 0.9491 和 0.6629。

表 4 扩展模型与传统模型的对比结果

| ID | Data Set | N# | M# | C# | 最优 $\alpha$ | 覆盖率 | | 测试精度 | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 传统模型 | 扩展模型 | 传统模型 | 扩展模型 |
| 1 | pima | 768 | 9 | 2 | 0.7 | 0.8279 | 1 | 0.6088 | 0.6445 |
| 2 | wine | 178 | 14 | 3 | 0.85 | 0.2222 | 0.5556 | 0.3982 | 0.4815 |
| 3 | climate | 540 | 21 | 2 | 0.6 | 0 | 1 | 0.5 | 0.9074 |
| 4 | breast | 569 | 32 | 2 | 0.75 | 0.3640 | 0.9561 | 0.5373 | 0.6864 |
| 5 | housing | 506 | 14 | 2 | 0.8 | 0.0542 | 0.9163 | 0.5172 | 0.8916 |
| 6 | parkinsons | 197 | 23 | 2 | 0.75 | 0 | 1 | 0.5 | 0.7949 |
| 7 | glass | 214 | 10 | 6 | 0.75 | 0.0116 | 0.9651 | 0.1764 | 0.1919 |
| 8 | SPECTF | 269 | 45 | 2 | 0.55 | 0.0185 | 1 | 0.4907 | 0.8657 |
| 9 | Hill_Valley | 606 | 101 | 2 | 0.8 | 0.9753 | 0.9959 | 0.5082 | 0.5104 |
| 10 | ionosphere | 351 | 34 | 2 | 0.9 | 0 | 1 | 0.5 | 0.8582 |
| 11 | Heart Disease | 87 | 14 | 5 | 0.9 | 0.4 | 1 | 0.1914 | 0.3429 |
| 12 | sonar | 208 | 60 | 2 | 0.75 | 0 | 1 | 0.5 | 0.7798 |
| 平均值 | | | | | | 0.2395 | **0.9491** | 0.4524 | **0.6629** |

注：表中 N# 代表样例个数，M# 代表属性个数，C# 代表类别个数。

需要说明的是，这里考虑到计算时间的问题，没有遍历所有 $\alpha$ 和 $\beta$ 的取值来求它们的最优取值，而是采用先固定一个再求另外一个最优值的方法，即表 4 中对于扩展模型的结果不一定是最优情况下的取值。这说明基于扩展优势关系的变精度粗糙集模型在性能上还有提升的空间。为了更加直观地说明参数的取值方法以及参数与模型性能的关系，随机选取了 3 组数据，实验结果如图 1 所示（星线表示 sonar 的实验结果，三角线表示 parkinsons 的实验结果，方块线表示 SPECTF 的实验结果）。

图 1(a)表示固定 $\beta=0.6$ 时，不同 $\alpha$ 下的覆盖率，其中横坐标是 $\alpha$ 的值，纵坐标是覆盖率。可以看到，每条线在 $\alpha<1$ 时都高于 $\alpha=1$ 的情况。也就是说，利用扩展优势关系下变精度粗糙集方法得到的覆盖率要比采用传统优势关系下变精度粗糙集方法得到的覆盖率高。数据集 sonar 上取阈值 $\alpha=[0.95,1]$ 时采用所提方法所得的覆盖率与传统优势关系方法得到的覆盖率相同。图 5(b)表示取 $\beta=0.6$ 时，不同 $\alpha$ 下的测试精度，其中横坐标是 $\alpha$ 的值，纵坐标是测试精度。可以看到，每条线的主体上都比 $\alpha=1$ 时的情况高，也就是说，利用扩展优势关系下变精度粗糙集方法得到的测试精度要比采用传统优势关系下变精度粗糙集方法得到的测试精度高，至少不低于后者。星线前后保持水平，$\alpha=0.75$ 时最高，即在 $\alpha=0.75$ 处取得最优值。

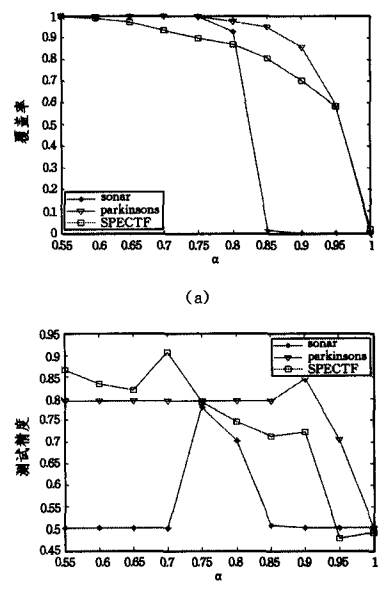


图 1 $\alpha$ 变化时覆盖率与测试精度的比较

可见，参数取过大和过小都会影响到模型的分类结果，相比传统模型，变精度模型具有更好的分类效果。

**结束语** 本文对优势关系进行了扩展，对基于扩展优势关系下的变精度粗糙集模型进行了详细的描述，主要目的在于放宽优势集的定义，提高传统优势关系粗糙集模型的泛化

能力以进一步满足实际需要，并对模型的可行性和有效性进行了算例和实验验证。在实验分析中，从 UCI 数据集中选取了几组数据集，说明了所提出的方法明显优于传统优势关系下的变精度粗糙集模型。另外，由表 4 观察到两类问题的测试精度较高，而多类问题如 glass 与 Heart Disease 的测试精度较低，可以进一步研究分类数目对测试精度的影响。不过与传统模型相比，所提出的扩展模型在这两组数据上仍能明显提高测试精度。从图 5 中发现 SPECTF 在 $\alpha=0.95$ 处的值小于 $\alpha=1$ 时的值，但是其它位置的值均比较大。未来可以进一步探讨类别数及参数的取值对测试精度的影响。

## 参 考 文 献


[1] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Sciences, 1982, 11(3): 341-356

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

[3] 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

[4] Ziarko W. Variable precision rough set model [J]. Journal of Computer and System Sciences, 1993, 46(1): 39-59

[5] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2003: 123-125

[6] Xie F, Lin Y, Ren W. Optimizing model for land use land cover retrieval from remote sensing imagery based on variable precision rough set[J]. Ecological Modeling, 2011, 222(2): 232-240

[7] Xie G, Yue W, Wang S, et al. Dynamic risk management in petroleum project investment based on a variable precision rough set model [J]. Technological Forecasting & Social Change, 2010, 77(6): 891-901

[8] Chen Hang, Tao Jun, Zhang Jian-de. Intrusion detection research based on variable precision rough set[J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2014, 35(2): 196-199 (in Chinese)
陈行，陶军，张建德. 基于变精度粗糙集的入侵检测研究[J]. 青岛科技大学学报(自然科学版)，2014, 35(2): 196-199

[9] Greco S, Matarazzo B, Slowinski R. Rough Approximation of a preference relation by dominance relations[J]. European Journal of Operational Research, 1999, 117(1): 63-83

[10] Greco S, Matarazzo B, Slowinski B. Rough approximation by dominance relations[J]. International Journal of Intelligent Systems, 2002, 17(2): 153-171

[11] Li S Y, Li T R. Incremental update of approximations in dominance-based rough sets approach under the variation of attribute values[J]. Information Sciences, 2015, 294: 348-361

[12] Li S. Y, Li T R, Liu D. Incremental updating approximations in dominance-based rough sets approach under the variation of attribute set[J]. Knowledge-Based Systems, 2013, 40(1): 17-26

[13] Du W, Hu B. Approximate distribution reducts in inconsistent interval-valued ordered decision tables[J]. Information Sciences, 2014, 271(7): 93-114

[14] Wong S K M, Ziarko W. Comparison of the probabilistic approximate classification and the fuzzy set model[J]. Fuzzy Sets and Systems, 1987, 21(3): 357-362

[15] Yao Y Y, Wong S K M. A decision theoretic framework for approximating concepts[J]. International Journal of Man -machine Studies, 1992, 37(6): 793-809

[16] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International journal of general system, 1990, 17(2): 191-209

[17] Parthaláin N M, Shen Q, Jensen R. A distance measure approach to exploring the rough set boundary region for attribute reduction[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(3): 306-317

[18] Hu Jie, Zhao Hui, Huang Chang-qiang, et al. The variable precision dominance rough set in the application of UCAV in threat estimation[J]. Journal of Air Force Engineering University (Natural Science Edition), 2009, 10(5): 27-31 (in Chinese)
胡杰，赵辉，黄长强，等. 优势变精度粗糙集在 UCAV 威胁估计中的应用[J]. 空军工程大学学报(自然科学版)，2009, 10(5): 27-31

[19] Luo Gong-zhi, Yang Xiao-jiang. Variable precision dominance rough set attribute reduction preferential algorithm[J]. Chinese Journal of Management Science, 2009, 17(2): 169-175 (in Chinese)
骆公志，杨晓江. 变精度优势粗糙集属性约简择优算法[J]. 中国管理科学，2009, 17(2): 169-175

[20] Inuiguchi M, Yoshioka Y, Kusunoki Y. Variable-precision dominance -based rough set approach and attribute reduction[J]. International Journal of Approximate Reasoning, 2009, 50(8): 1199-1214

[21] Karami J, Alimohammadi A, Seifouri T. Water quality analysis using a variable consistency dominance-based rough set approach[J]. Computers, Environment and Urban Systems, 2014, 43(1): 25-33

[22] Zhai L, Khoo L, Zhong Z. A rough set based decision support approach to improving consumer affective satisfaction in product design[J]. International Journal of Industrial Ergonomics, 2009, 39(2): 295-302

[23] 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


(上接第 226 页)


[11] Li Hai-feng, Li Chuan-guo. Note on deep architecture and deep learning algorithms [J]. Journal of Heibei University (Natural Science Edition), 2012, 32(5): 538-544 (in Chinese)
李海峰，李纯果. 深度学习结构和算法比较分析[J]. 河北大学学报(自然科学版)，2012, 32(5): 538-544

[12] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504-507

[13] Angluin D, Laird P. Learning from noisy examples[J]. Machine Learning, 1988, 2(4): 343-370

[14] CMU face images data set [OL]. http://archive.ics.uci.edu/ml/datasets

[15] Kanade T, Cohn J, Tian Y. Comprehensive Database for Facial Expression Analysis[C]// Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition. EE Computer Society, 2000: 46-53

[16] Bashyal S, Venayagamoorthy G K. Recognition of facial expressions using Gabor wavelets and learning vector quantization [J]. Engineering Applications of Artificial Intelligence, 2008, 1(7): 1056-1064

[17] Ojala T, Pietikainan M. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987

[18] Bashyal S, Venayagamoorthy G K. Recognition of Facial Expressions Using Gabor Wavelets and Learning Vector Quantization [J]. Engineering Applications of Artificial Intelligence, 2008, 21(7): 1056-1064