基于差异关系的变精度粗糙集知识约简算法研究

焦 娜

(华东政法大学信息科学与技术系 上海 201620)

摘 要 有效的知识约简算法是粗糙集理论的重要研究内容。粗糙集是一个去掉冗余特征的有效工具。经典的粗糙集方法要求数值用离散数据表达,对于连续值则在处理前必须进行离散化处理。真实数据往往存在连续值,为了避免运用粗糙集方法所必需的离散化过程带来的信息丢失,将差异关系应用于粗糙集的知识约简。为进一步增强差异关系粗糙集对噪声数据的适应能力,提出基于差异关系的变精度粗糙集知识约简算法,并分析差异关系下变精度粗糙集模型参数的特性,给出依赖度和参数范围关系描述,将参数取值从点扩展到区间范围。在UCI数据库的数据集上进行实验,结果证明了所提方法及相关理论的有效性。

关键词 粗糙集理论,差异关系,变精度,参数范围,属性依赖度

中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.053

Research on Knowledge Reduction Algorithm Based on Variable Precision Tolerance Rough Set Theory

JIAO Na

(Department of Information Science and Technology, East China University of Political Science and Law, Shanghai 201620, China)

Abstract Knowledge reduction is an important research issue in rough set theory. Rough set theory is an efficient mathematical tool for further reducing redundancy. The main limitation of traditional rough set theory is the lack of effective methods for dealing with real-valued data. However, practical data sets are always continuous. This has been addressed by employing discretization methods, which may result in information loss. This paper investigated one approach combining tolerance relation together with rough set theory. In order to enhance the ability to adapt to the noise data, this paper explored the knowledge reduction algorithm based on variable precision tolerance rough set theory. The characteristics of parameter were analyzed. The relationship between the classification quality and parameter interval was described, and the parameter value was extended to interval range. The experimental results demonstrate that our proposed algorithm and the related theory are effective.

Keywords Rough set theory, Tolerance relation, Variable precision, Parameter interval, Degree of dependency of feature

1 引言

自 Pawlak[1] 在 20 世纪 80 年代初提出粗糙集理论(Rough Set Theory)以来,该理论在理论模型、算法构造和系统开发上都取得了显著的成果。知识约简是粗糙集理论的重要研究内容之一。经典的粗糙集理论模型主要是通过不可分辨关系(等价关系)得到相对于论域的下近似、上近似、正区域、负区域及边界域,最终得到知识约简。

经典粗糙集模型运用等价关系计算知识约简时,对处理的数据要求较为严格,要求数据是离散的数值(如整数、字符串型、枚举型),真正获取的数据往往存在连续值,对于连续值的数据则必须进行离散化[18],但离散化过程必定会造成信息某种程度的损失。因此,本文引入差异度,用属性子集上的对象间的差异来度量其属于同一类的可能性,并定义了差异关系和差异关系下的属性依赖度,利用差异关系粗糙集对连续值的属性进行处理避免了离散化过程所造成的信息损失,能最大限度地保持数据集的分类能力,使粗糙集理论只能处理离散化数据的问题得到了解决。

经典粗糙集是在理想状态下得到的知识约简结果,没有考虑误差的存在。实际问题中采集到的数据往往存在各种误差,差异关系下的粗糙集理论在实际应用中会面临噪声数据,如果考虑噪声数据,则得到的约简不够理想或者新对象的预测能力会降低。很多研究者对此进行了讨论[2-9,20,21],Ziarko提出了变精度粗糙集模型,该模型给出了错误率低于预先给定值的处理策略,定义了变精度下的正区域、负区域和边界域,讨论了变精度粗糙集模型的有关性质[2];Katzberg 和 Ziarko 进一步提出了不对称边界变精度粗糙集模型,使此模型更加一般化,从而拓广了变精度粗糙集模型的应用范围[3];米据生和吴伟志等人讨论了变精度粗糙集模型,并提出知识约简方法[4];张贤勇等人定义了变精度粗糙集的近似算子,得到较好的结论[5,6];Yao[8,9]等讨论了概率粗糙集模型,并取得较好的理论模型。为进一步增强差异关系粗糙集对噪声数据的适应能力,本文将变精度引入差异关系的粗糙集模型中,提出基于差异关系的变精度粗糙集知识约简算法,并分析差异关系下变精度粗糙集模型参数的特性,给出依赖度和参数范围关系描述。在UCI数据库的数据集上进行实验,结果证明了

到稿日期:2014-06-14 返修日期:2014-09-16 本文受国家社科基金青年项目(13CFX049),上海高校青年教师培养资助计划(hdzf10008)资助。

焦 娜(1977—),女,博士,讲师,主要研究方向为人工智能、模式识别、粗糙集、数据挖掘等,E-mail:zdx.jn@163.com。

本文所提方法及相关理论的有效性。

2 粗糙集理论

2.1 基本概念

下面简要介绍经典粗糙集的基本概念[10-14]。

定义 1 $DT=(U,C\cup D,V,f)$是一个决策表，其中U为非空有限论域；C、D分别为条件属性集和决策属性集；$V=\bigcup_{a\in C\cup D}V_a$为属性值域；$f:U\times(C\cup D)\to V$为信息函数。每个属性子集$B\subseteq C\cup D$决定了一个不可分辨关系$IND(B)$：

$$IND(B)=\{(x,y)\in U\times U\mid \forall a\in B, f(x,a)=f(y,a)\}$$

显然，$IND(B)$是一个等价关系。论域U关于不可分辨关系B的划分$U/IND(B)$定义为：

$$U/IND(B)=\{[x]_B:x\in U\}$$

$[x]_B$表示U中所有与x在关系$IND(B)$下等价的元素构成的集合$[x]_B=\{y\mid \forall a\in B, f(x,a)=f(y,a)\}$，即等价类。

定义 2 在决策表$DT=(U,C\cup D,V,f)$中，$X\subseteq U$关于$B\subseteq C$的下近似集和上近似集分别记为$\underline{B}X$、$\overline{B}X$：

$$\underline{B}X=\bigcup\{[x]_B\in U/IND(B)\mid [x]_B\subseteq X\}$$

$$\overline{B}X=\bigcup\{[x]_B\in U/IND(B)\mid [x]_B\cap X\neq\emptyset\}$$

定义 3 在决策表$DT=(U,C\cup D,V,f)$中，$X\subseteq U$，决策属性集D相对条件属性子集$B\subseteq C$的正区域、负区域和边界域分别定义为：

$$POS_B(D)=\bigcup_{X\in U/IND(D)}\underline{B}X$$

$$NEG_B(D)=U-\bigcup_{X\in U/IND(D)}\overline{B}X$$

$$BND_B(D)=\bigcup_{X\in U/IND(D)}\overline{B}X-\bigcup_{X\in U/IND(D)}\underline{B}X$$

定义 4 在决策表$DT=(U,C\cup D,V,f)$中，决策属性集D相对条件属性子集$B\subseteq C$的依赖度定义为：

$$\gamma_B(D)=|POS_B(D)|/|U|$$

3 差异关系的粗糙集

在经典粗糙集中，等价关系只能对离散化的数据进行处理，对于连续值的数据，离散化过程会带来信息丢失。为此，本文定义了差异度、差异关系及相对应的正区域等概念，用差异关系来代替等价关系，从而减少了信息损失。

3.1 差异关系

定义 5[15-17] 在决策表$DT=(U,C\cup D,V,f)$中，$a\in C\cup D$，$x,y\in U$，属性a对于任意两个对象x、y间的差异度定义为：

$$F_a(x,y)=\frac{|a(x)-a(y)|}{|a_{\max}-a_{\min}|}$$

$a_{\max}$、$a_{\min}$分别表示属性a的最大值和最小值。$F_a(x,y)$越小表示两个对象的差异越小，属于同一类的可能性越大；反之，属于同一类的可能性越小。对于属性子集，对象间综合差异度有如下定义。

定义 6 在决策表$DT=(U,C\cup D,V,f)$中，$B\subseteq C\cup D$，$x,y\in U$，任意两个对象x、y关于属性子集B的综合差异度定义为：

$$F_{B,\tau}(x,y)=\{(x,y)\mid \frac{\sum_{a\in B}F_a(x,y)}{|B|}\leqslant\tau\}$$

其中，$\tau\in(0,0.5]$是一个阈值。

定义 7 在决策表$DT=(U,C\cup D,V,f)$中，$B\subseteq C\cup D$，$\tau\in(0,0.5]$，论域U关于属性子集B的τ差异关系定义为：

$$F_{B,\tau}=\{(x,y)\in U\times U\mid \forall a\in B,(x,y)\in F_{B,\tau}(x,y)\}$$

论域U关于差异关系$F_{B,\tau}$的划分用$U/F_{B,\tau}$表示。

$$U/F_{B,\tau}=\{F_{B,\tau}(x):x\in U\}$$

$F_{B,\tau}(x)$表示U中所有与x在差异关系$F_{B,\tau}$下的元素构成的集合$F_{B,\tau}(x)=\{y\mid \forall a\in B,(x,y)\in F_{B,\tau}(x,y)\}$，即差异类。

定义 8 在决策表$DT=(U,C\cup D,V,f)$中，$X\subseteq U$，$B\subseteq C$，$\tau\in(0,0.5]$，X关于B的τ下近似集和上近似集分别记为$\underline{B_\tau}X$、$\overline{B_\tau}X$：

$$\underline{B_\tau}X=\{x\mid F_{B,\tau}(x)\subseteq X\}$$

$$\overline{B_\tau}X=\{x\mid F_{B,\tau}(x)\cap X\neq\emptyset\}$$

$\langle\underline{B_\tau}X,\overline{B_\tau}X\rangle$为差异粗糙集。

定义 9 在决策表$DT=(U,C\cup D,V,f)$中，$X\subseteq U$，$B\subseteq C$，$\tau\in(0,0.5]$，决策属性集D相对条件属性子集B的τ正区域、负区域和边界域分别定义为：

$$POS_{B,\tau}(D)=\bigcup_{X\in U/F_{D,\tau}}\underline{B_\tau}X$$

$$NEG_{B,\tau}(D)=U-\bigcup_{X\in U/F_{D,\tau}}\overline{B_\tau}X$$

$$BND_{B,\tau}(D)=\bigcup_{X\in U/F_{D,\tau}}\overline{B_\tau}X-\bigcup_{X\in U/F_{D,\tau}}\underline{B_\tau}X$$

定义 10 在决策表$DT=(U,C\cup D,V,f)$中，$B\subseteq C$，$\tau\in(0,0.5]$，决策属性集D相对条件属性子集B的τ依赖度定义为：

$$\gamma_{B,\tau}(D)=|POS_{B,\tau}(D)|/|U|$$

差异粗糙集属性约简方法是通过$\gamma_{B,\tau}(D)$来度量属性子集的重要性。

4 基于差异关系的变精度粗糙集知识约简算法

实际应用中，噪音数据有时是无法避免的，要满足经典粗糙集理论下近似、上近似的严格条件是非常困难的。为解决这个矛盾，允许噪音数据或误差的存在，也不影响对数据处理的效果。本文将变精度粗糙集模型应用在差异关系粗糙集模型中，定义了差异关系下的变精度粗糙集模型，提出了基于差异关系的变精度粗糙集知识约简算法。分析参数的特性，给出依赖度和参数范围关系描述。

4.1 差异关系下的变精度粗糙集模型

定义 11 在决策表$DT=\langle U,C\cup\{d\},V,f\rangle$中，$\tau\in(0,0.5]$，$X_i\in U/F_{C,\tau}(i=1,2,\cdots,|U/F_{C,\tau}|)$，$Y_j\in U/F_{D,\tau}(j=1,2,\cdots,|U/F_{D,\tau}|)$，$B\subseteq C$，$Y_j$关于$B$的$\beta$-下近似和$\beta$-上近似分别记为$\underline{B_\tau^\beta}Y_j$、$\overline{B_\tau^\beta}Y_j$：

$$\underline{B_\tau^\beta}Y_j=\bigcup\{X_i\mid \frac{|X_i\cap Y_j|}{|X_i|}\geqslant\beta, X_i\in U/F_{B,\tau}\}$$

$$\overline{B_\tau^\beta}Y_j=\bigcup\{X_i\mid \frac{|X_i\cap Y_j|}{|X_i|}>1-\beta, X_i\in U/F_{B,\tau}\}$$

其中，$0.5<\beta\leqslant 1$。

定义 12 在决策表$DT=\langle U,C\cup\{d\},V,f\rangle$中，$\tau\in(0,0.5]$，$\beta\in(0.5,1]$，$X_i\in U/F_{C,\tau}(i=1,2,\cdots,|U/F_{C,\tau}|)$，$Y_j\in U/F_{D,\tau}(j=1,2,\cdots,|U/F_{D,\tau}|)$，决策属性集$D$相对条件属性子集$B\subseteq C$的$\tau$的$\beta$-正区域、$\beta$-负区域和$\beta$-边界域分别定义为：

$$POS_{B,\tau}^\beta(D)=\bigcup_{Y_j\in U/F_{D,\tau}}\underline{B_\tau^\beta}Y_j$$

$$NEG^{\beta}_{B,\tau}(D)=U-\bigcup_{Y_j\in U/F_{D,\tau}}\overline{B^{\beta}_{\tau}}Y_j$$

$$BN^{\beta}_{B,\tau}(D)=\bigcup_{Y_j\in U/F_{D,\tau}}\overline{B^{\beta}_{\tau}}Y_j-\bigcup_{Y_j\in U/F_{D,\tau}}\underline{B^{\beta}_{\tau}}Y_j$$

定义 13 在决策表 $DT=(U,C\cup D,V,f)$ 中，$B\subseteq C$，$\tau\in(0,0.5]$，$\beta\in(0.5,1]$，决策属性集 D 相对条件属性子集 B 的 τ 的 β-依赖度定义为：

$$\gamma^{\beta}_{B,\tau}(D)=|POS^{\beta}_{B,\tau}(D)|/|U|$$

4.2 基于差异关系的变精度粗糙集知识约简算法

基于差异关系的变精度粗糙集知识约简算法(DDTRS)的具体步骤描述如下。

步骤 1 设 τ 初值，根据定义 5－定义 7 计算决策表的差异关系及差异关系下的划分。

步骤 2 设 β 初值，根据定义 11－定义 13 计算决策表在差异关系下的 β -下近似和 β -上近似，再计算决策属性集 D 相对条件属性集 C 的 τ 的 β -正区域、β -负区域、β -边界域和 β -依赖度。

步骤 3 从所有条件属性集合开始进行约简 $Red=C$。

步骤 4 对条件属性集合中的每个属性进行计算，判断去掉该属性后的 τ 的 β-依赖度和决策属性集 D 相对条件属性集 C 的 τ 的 β-依赖度是否相同。

i)若相同，$\gamma^{\beta}_{P-\{a\},\tau}(D)==\gamma^{\beta}_{C,\tau}(D)$，则继续执行步骤 5。

ii)若不同，$\gamma^{\beta}_{P-\{a\},\tau}(D)\neq\gamma^{\beta}_{C,\tau}(D)$，则转到步骤 6。

步骤 5 该属性为冗余属性，可以去掉该属性 $Red=Red-\{a\}$。

步骤 6 循环结束，输出约简结果 Red。

4.3 基于差异关系的变精度粗糙集知识约简算法的参数分析

定义 14 在决策表 $DT=(U,C\cup D,V,f)$ 中，$X_i\in U/F_{C,\tau}$ $(i=1,2,\cdots,|U/F_{C,\tau}|)$，$Y_j\in U/F_{D,\tau}(j=1,2,\cdots,|U/F_{D,\tau}|)$，集合 X_i 相对于集合 Y_j 的包含度量定义为：

$$ID(X_i,Y_j)=\begin{cases}\dfrac{|X_i\cap Y_j|}{|X_i|}, & \text{if } |X_i|>0\\ 0, & \text{if } |X_i|=0\end{cases}$$

定义 15 在决策表 $DT=(U,C\cup D,V,f)$ 中，$X_i\in U/F_{C,\tau}$ $(i=1,2,\cdots,|U/F_{C,\tau}|)$，$Y_j\in U/F_{D,\tau}(j=1,2,\cdots,|U/F_{D,\tau}|)$，集合 X_i 相对于集合 $U/F_{D,\tau}$ 的参数分界点定义为：

$$\alpha_i=\mathrm{Max}(ID(X_i,Y_j)),j=1,2,\cdots,|U/F_{D,\tau}|$$

5 实例分析

为说明本文提出的基于差异关系的变精度粗糙集知识约简算法的有效性，本文对表 1 所列的决策表进行约简。表 1 中共 10 个对象，3 个条件属性，1 个决策属性，表中的属性值都是连续的。设 $\tau=0.2$，$C=\{E,F,G\}$，$D=\{H\}$。

表 1 数据集

对象	E	F	G	H
o_1	0.3	90	3	53
o_2	0.3	90	4	14
o_3	2	11	3	14
o_4	2	40	3	14
o_5	2.2	41	4	13
o_6	2.3	132	18	14
o_7	2.1	133	19	56
o_8	2.2	134	19	52
o_9	5	134	18	55
o_{10}	5	131	19	53

差异类分别是：

$$U/F_{D,\tau}=\{\{o_2,o_3,o_4,o_5,o_6\},\{o_1,o_7,o_8,o_9,o_{10}\}\}$$

$$U/F_{C,\tau}=\{\{o_1,o_2\},\{o_3\},\{o_4,o_5\},\{o_6,o_7,o_8\},\{o_9,o_{10}\}\}$$

$$U/F_{C-\{E\},\tau}=U/F_{\{F,G\},\tau}=\{\{o_1,o_2\},\{o_3\},\{o_4,o_5\},\{o_6,o_7,o_8,o_9,o_{10}\}\}$$

$$U/F_{C-\{F\},\tau}=U/F_{\{E,G\},\tau}=\{\{o_1,o_2\},\{o_3,o_4,o_5\},\{o_6,o_7,o_8\},\{o_9,o_{10}\}\}$$

$$U/F_{C-\{G\},\tau}=U/F_{\{E,F\},\tau}=\{\{o_1,o_2\},\{o_3\},\{o_4,o_5\},\{o_6,o_7,o_8\},\{o_9,o_{10}\}\}$$

$$U/F_{C-\{E,F\},\tau}=U/F_{\{G\},\tau}=\{\{o_1,o_2,o_3,o_4,o_5\},\{o_6,o_7,o_8,o_9,o_{10}\}\}$$

$$U/F_{C-\{E,G\},\tau}=U/F_{\{F\},\tau}=\{\{o_1,o_2\},\{o_3\},\{o_4,o_5\},\{o_6,o_7,o_8,o_9,o_{10}\}\}$$

$$U/F_{C-\{E,G\},\tau}=U/F_{\{F\},\tau}=\{\{o_1,o_2\},\{o_3,o_4,o_5,o_6,o_7,o_8\},\{o_9,o_{10}\}\}$$

DDTRS 方法得到的约简结果：

设 $\beta=0.8$，相对于条件属性集 C 的 τ 的 β-下近似为：

$$\underline{C^{\beta}_{\tau}}\{o_2,o_3,o_4,o_5,o_6\}=\underline{\{E,F,G\}^{\beta}_{\tau}}\{o_2,o_3,o_4,o_5,o_6\}=\{o_3,o_4,o_5\}$$

$$\underline{C^{\beta}_{\tau}}\{o_1,o_7,o_8,o_9,o_{10}\}=\underline{\{E,F,G\}^{\beta}_{\tau}}\{o_1,o_7,o_8,o_9,o_{10}\}=\{o_9,o_{10}\}$$

因此，决策属性集 D 相对条件属性集 C 的 τ 的 β-正区域为：

$$POS^{\beta}_{C,\tau}(D)=\bigcup^{\beta}_{Y_j\in U/F_{D,\tau}}\underline{C^{\beta}_{\tau}}Y_j=\underline{C^{\beta}_{\tau}}\{o_2,o_3,o_4,o_5,o_6\}\cup\underline{C^{\beta}_{\tau}}\{o_1,o_7,o_8,o_9,o_{10}\}=\{o_3,o_4,o_5,o_9,o_{10}\}$$

决策属性集 D 相对条件属性集 C 的 τ 的 β-依赖度为：

$$\gamma^{\beta}_{C,\tau}(D)=\frac{|POS^{\beta}_{C,\tau}(D)|}{|U|}=\frac{|\{o_3,o_4,o_5,o_9,o_{10}\}|}{|\{o_1,o_2,o_3,o_4,o_5,o_6,o_7,o_8,o_9,o_{10}\}|}=\frac{5}{10}$$

对于属性集 $C-\{G\}$，决策属性集 D 相对条件属性集 $C-\{G\}$ 的 τ 的 β-依赖度为：

$$\gamma^{\beta}_{C-\{G\},\tau}(D)=\frac{|POS^{\beta}_{C-\{G\},\tau}(D)|}{|U|}=\frac{|\{o_3,o_4,o_5,o_9,o_{10}\}|}{|\{o_1,o_2,o_3,o_4,o_5,o_6,o_7,o_8,o_9,o_{10}\}|}=\frac{5}{10}$$

$$\gamma^{\beta}_{C-\{G\},\tau}(D)=\gamma^{\beta}_{\{E,F\},\tau}(D)=\gamma^{\beta}_{C,\tau}(D)=\frac{5}{10}$$

因此，可以在属性集 C 中删除属性 G。类似地，决策属性集 D 相对条件属性集 $\{E,F\}-\{E\}$ 的 τ 的 β-依赖度为：

$$\gamma^{\beta}_{\{E,F\}-\{E\},\tau}(D)=\frac{|POS^{\beta}_{\{E,F\}-\{E\},\tau}(D)|}{|U|}=\frac{|\{o_1,o_2,o_4,o_6,o_7,o_8,o_9,o_{10}\}|}{|\{o_1,o_2,o_3,o_4,o_5,o_6,o_7,o_8,o_9,o_{10}\}|}=\frac{8}{10}$$

$$\gamma^{\beta}_{\{E,F\}-\{E\},\tau}(D)=\frac{8}{10}\neq\gamma^{\beta}_{C,\tau}(D)=\frac{5}{10}$$

最后，算法终止，DDTRS 方法得到的约简结果为 $\{E,F\}$。

DDTRS 方法的参数分析：

对条件属性集 C，设 $X_1=\{o_3\},X_2=\{o_4,o_5\},X_3=\{o_1,o_2\},X_4=\{o_6,o_7,o_8\},X_5=\{o_9,o_{10}\}$；$Y_1=\{o_2,o_3,o_4,o_5,o_6\}$，$Y_2=\{o_1,o_7,o_8,o_9,o_{10}\}$，$U/F_{D,\tau}=\{Y_1,Y_2\}=\{\{o_2,o_3,o_4,o_5,o_6\},\{o_1,o_7,o_8,o_9,o_{10}\}\}$；$ID(X_1,Y_1)=\dfrac{|X_1\cap Y_1|}{|X_1|}=1$，$ID$

$(X_1,Y_2)=\frac{|X_1\cap Y_2|}{|X_1|}=0$。所以，$X_1$ 相对于 $U/F_{D,\tau}$ 的分界点为 $\alpha_1=\mathrm{Max}(ID(X_1,Y_j))=1(j=1,2)$，当 $\beta\in(0.5,1]$ 时，Y_j $(j=1,2)$ 相对于 X_1 的 β-下近似为 $X_1=\{o_3\}$；同理，X_2 相对于 $U/F_{D,\tau}$ 的分界点为 $\alpha_2=1$，当 $\beta\in(0.5,1]$ 时，Y_j 相对于 X_2 的 β-下近似为 $\{o_4,o_5\}$；X_3 相对于 $U/F_{D,\tau}$ 的分界点为 $\alpha_3=0.5$，当 $\beta=0.5$ 时，Y_j 相对于 X_3 的 β-下近似为 $\{o_1,o_2\}$；X_4 相对于 $U/F_{D,\tau}$ 的分界点为 $\alpha_4=0.667$，当 $\beta\in(0.5,0.667]$ 时，Y_j 相对于 X_4 的 β-下近似为 $\{o_6,o_7,o_8\}$；X_5 相对于 $U/F_{D,\tau}$ 的分界点为 $\alpha_5=1$，当 $\beta\in(0.5,1]$ 时，Y_j 相对于 X_5 的 β-下近似为 $\{o_9,o_{10}\}$。

当 $\beta\in(0.5,0.667]$ 时，$POS^{\beta}_{C,\tau}(D)=\bigcup_{Y_j\in U/F_{D,\tau}}\underline{C^{\beta}_{\tau}}Y_j=\{o_3\}\cup\{o_4,o_5\}\cup\{o_6,o_7,o_8\}\cup\{o_9,o_{10}\}=\{o_3,o_4,o_5,o_6,o_7,o_8,o_9,o_{10}\}$，$\gamma^{\beta}_{C,\tau}(D)=\frac{|POS^{\beta}_{C,\tau}(D)|}{|U|}=\frac{8}{10}$；当 $\beta\in(0.667,1]$ 时，$POS^{\beta}_{C,\tau}(D)=\{o_3,o_4,o_5,o_9,o_{10}\}$，$\gamma^{\beta}_{C,\tau}(D)=\frac{5}{10}$。

对决策表表 1 进行差异关系下的变精度粗糙集知识约简算法时，针对不同的 β 值，得到的不同的结果，具体结果见表 2。

表 2　不同 β 范围得到的不同约简结果

参数范围	依赖度	约简结果
$\beta\in(0.5,0.667]$	$\frac{8}{10}$	{E}或{F}
$\beta\in(0.667,0.8]$	$\frac{5}{10}$	{E,F}
$\beta\in(0.8,1]$	$\frac{5}{10}$	{F}

从表 2 得到，在参数 β 变化过程中，属性依赖度和约简结果也都在变化，当 $\beta\in(0.5,0.667]$ 时，约简结果是 $\{E\}$ 或 $\{F\}$；当 $\beta\in(0.8,1]$ 时，约简结果也是 $\{F\}$，但是属性依赖度不同。当 $\beta\in(0.667,0.8]$ 和 $\beta\in(0.8,1]$ 时，约简结果虽然不同，但属性依赖度都是 $\frac{5}{10}$。由于参数范围的不同，从决策表中得到的信息量在变化，属性依赖度和约简结果也在变化。

6　实验结果与分析

选用了 UCI 机器学习数据库中的 Post-operative 数据集来验证本文提出的算法。Post-operative 数据集包含有 L-Core、L-Surf、L-O_2、L-Bp、Surf-Stbl、Bp-Stbl、Core-Stbl、Comfort 8 个条件属性和 Adm-Decs 1 个决策属性，其中，Surf-Stbl、Bp-Stbl、Core-Stbl 3 个条件属性和 Adm-Decs 1 个决策属性的数值为离散值，其他 5 个条件属性的数值为连续值。实验在 Intel Dual E2140 2.0GHz(处理器)，2G(内存)，Window XP(操作系统)的个人计算机上进行。实验平台是 VC++6.0 及 SQL server 数据库。

Post-operative 数据集是一个不完备的数据集，用文献[19]方法对有缺失值的数据进行完备化。对于 DDTRS 算法，首先设 $\tau=0.2$。对 Post-operative 数据集计算差异类，得到 70 个条件差异类和 3 个决策差异类。不同的参数范围对应不同的属性依赖度和约简结果，具体结果见表 3。分别用 m_1 表示 L-Core、m_2 表示 L-Surf、m_3 表示 L-O_2、m_4 表示 L-Bp、m_5 表示 Surf-Stbl、m_6 表示 Bp-Stbl、m_7 表示 Core-Stbl、m_8 表示 Comfort。

表 3　不同参数范围的约简结果

参数范围	依赖度	约简结果
$\beta\in(0.5,0.571]$	$\frac{78}{90}$	$\{m_2,m_4,m_7\}$或$\{m_1,m_5,m_6,m_8\}$或$\{m_4,m_5,m_6,m_7\}$
$\beta\in(0.571,0.583]$	$\frac{78}{90}$	$\{m_2,m_3,m_5,m_6\}$
$\beta\in(0.583,0.6]$	$\frac{78}{90}$	$\{m_1,m_3,m_4,m_7\}$或$\{m_3,m_5,m_6,m_7,m_8\}$或$\{m_1,m_2,m_3,m_5,m_7,m_8\}$
$\beta\in(0.6,0.625]$	$\frac{78}{90}$	$\{m_4,m_8\}$或$\{m_3,m_6,m_8\}$或$\{m_1,m_3,m_4,m_5\}$
$\beta\in(0.625,0.667]$	$\frac{78}{90}$	$\{m_1,m_3,m_8\}$或$\{m_4,m_6,m_7\}$
$\beta\in(0.667,0.714]$	$\frac{72}{90}$	$\{m_1,m_3,m_8\}$
$\beta\in(0.714,1]$	$\frac{72}{90}$	$\{m_1,m_2,m_3,m_4,m_5,m_7,m_8\}$

参数的分界点分别是 0.571、0.583、0.6、0.625、0.667 和 0.714，随着参数范围的变化，得到的属性依赖度和约简结果都会相应变化。

结束语　经典粗糙集方法需要对连续值进行离散化处理，而离散化过程必定带来信息丢失，因此本文定义了差异关系，求得差异关系下的划分；又引入变精度，提出基于差异关系的变精度粗糙集知识约简算法，定义了参数分界点，分析了相关性质，讨论了参数在不同的区间范围时的属性依赖度及约简的变化。通过一个实例来说明本文所提方法，并在 UCI 数据库的数据集上进行实验，实验结果表明了本文所提知识约简算法及相关理论的有效性。

参 考 文 献

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

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

[3] Katzberg J D, Ziarko W. Variable precision rough sets with asymmetric bounds[C] // Ziarko W. ed. Proceedings of Rough Sets, and Fuzzy Sets and Knowledge Discovery (RSKD'93). London: Springer-Verlag, 1994: 167-176

[4] Mi J S, Wu W Z, Zhang W X. Approaches to knowledge reduction based on variable precision rough set model[J]. Information Sciences, 2004, 159(3): 255-272

[5] 张贤勇，莫智文. 变精度粗糙集[J]. 模式识别与人工智能，2004，17(2)：151-155

[6] Zhang X Y, Mo Z W, Xiong F, et al. Comparative study of variable precision rough set model and graded rough set model[J]. International Journal of Approximate Reasoning, 2012, 53(1): 104-116

[7] Zhang H Y, Leung Y, Zhou L. Variable-precision-dominance-based rough set approach to interval-valued information systems [J]. Information Sciences, 2013, 244(20): 75-272

[8] Yao Y Y. Probabilistic rough set approximations[J]. International Journal of Approximate Reasoning, 2008, 49(2): 255-271

[9] Yao Y Y, Yao B X. Covering based rough set approximations [J]. Information Sciences, 2012, 200: 91-107

[10] 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999，36(6)：681-684

[11] Yang X B, Xie J, Song X N, et al. Credible rules in incomplete decision system based on descriptors[J]. Knowledge-Based Systems, 2009, 22: 8-17

[12] Yu Y, Pedrycz W, Miao D Q. Neighborhood rough sets based multi-label classification for automatic image annotation[J]. Journal of Approximate Reasoning, 2013, 54(9): 1373-1387
[13] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001
[14] 苗夺谦. 粗糙集理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302
[15] Jensen R, Shen Q. Tolerance-based and fuzzy-rough feature selection[C]// Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ- IEEE'07). 2007: 877-882
[16] Parthaláin N M, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection[J]. Pattern Recognition, 2009, 42: 655-667
[17] Shen Q, Chouchoulas A. A rough-fuzzy approach for generating classification rules[J]. Pattern Recognition, 2002, 5: 2425-2438
[18] Grzymala-Busse J W. Discretization of numerical attributes[M]// Klösgen W, Zytkow J, eds. Handbook of Data Mining and Knowledge Discovery. Oxford University Press, 2002: 218-225
[19] Grzymala-Busse J W, Grzymala-Busse W J. Handling missing attribute values[M]// Maimon O, Rokach L, eds. Handbook of Data Mining and Knowledge Discovery. 2005: 37-57
[20] Min F, Zhu W. Attribute reduction of data with error ranges and test costs[J]. Information Sciences, 2012, 211: 48-67
[21] Zhao H, Min F, Zhu W. Cost-Sensitive Feature Selection of Numeric Data with Measurement Errors[J]. Journal of Applied Mathematics, 2013, 2013

(上接第 254 页)

EDA, EDSFLA 求解该问题时 CPU 耗时最少，且随着各服务节点候选服务数量规模的增加，CPU 开销仅呈现线性增加趋势，能够有效地满足大部分服务动态优化选择问题的求解。

5.3 算法收敛性能对比

本实验目的是在候选服务数量固定时，对 3 种服务选择算法的收敛性能进行对比。实验中为每个节点生成 100 个候选服务实例，实验重复 10 次取平均结果，图 7 给出了 3 个算法的进化曲线。

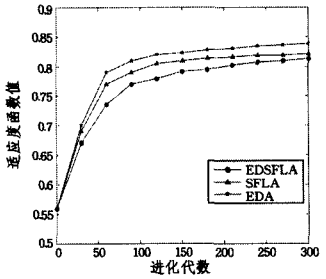

图 7 算法收敛性对比

可以看出，与 SFLA 和 EDA 相比，EDSFLA 的收敛速度最快且收敛精度更高，具有更强的全局寻优能力，这说明算法在借鉴分布估计进化思想的基础上所进行的改进是有效的。

结束语 为了更好地解决云计算环境 QoS 全局最优服务动态选择问题，本文提出了基于分布估计蛙跳算法的服务动态选择方法。该方法采用遗传算法的交叉操作重新设计了蛙跳算法的进化算子，利用分布估计进化策略改进标准蛙跳算法的青蛙更新模式，增强了各子群之间的相互学习能力，能够有效避免陷入局部最优。通过仿真实验验证了该方法的可行性和有效性，与标准蛙跳算法和分布估计算法相比，在求解过程中，本文提出的算法具有更优的收敛性能。

参 考 文 献

[1] Buyya R, Yeo C S, Venugopal S, et al. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility [J]. Future Generation Computer Systems, 2012, 25(6): 599-616
[2] 罗军舟, 金嘉晖, 宋爱波, 等. 云计算: 体系架构与关键技术[J]. 通信学报, 2011, 32(7): 3-21
[3] Majithia S, Walker D W, Gray W A. A Framework for Automated Service Composition in Service-oriented Architectures[C]// ESWS 2008 Congress. Berlin, Germany: Springer Verlag, 2012: 269-283
[4] Yu Tao, Lin K J. Service Selection Algorithms for Composing Complex Services with Multiple QoS Constraints[C]// The 3rd International Conference on Service Oriented Computing Congress. Amsterdam, Holland: Springer Verlag, 2010: 130-143
[5] 董元元, 倪宏, 邓浩江, 等. Qos 全局最优的服务选择策略[J]. 中南大学学报: 自然科学版, 2013, 42(10): 3086-3094
[6] 刘旋, 廖明潮. 基于人工鱼群算法的 QoS 全局最优 Web 服务最优 Web 服务选择的研究[J]. 计算机应用与软件, 2013, 30(8): 87-90
[7] 孙黎阳, 林剑柠, 毛少杰. 基于改进粒子群优化算法的网络化仿真任务共同体服务选择[J]. 兵工学报, 2012, 33(11): 1393-1403
[8] Elbeltagi E, Hegazy Grierson D. A modified shuffled frog-leaping optimization algorithm. Applications to project management [J]. Structure and Infrastructure Engineering, 2012, 3(1): 53-60
[9] Antariksha B. A clonal selection based shuffled frog leaping algorithm[C]// IEEE Advance Computing Congress. New York: IEEE, 2013: 125-130
[10] 张恒巍, 卫波. 基于分布估计蛙跳算法的云资源调度方法[J]. 计算机应用研究, 2014, 31(10): 30-34
[11] 罗雪晖, 杨烨, 李霞. 改进混合蛙跳算法求解旅行商问题[J]. 通信学报, 2013, 30(7): 130-135
[12] 王尚广, 孙其博, 杨放春. 基于全局 QoS 约束分解的 Web 服务动态选择[J]. 软件学报, 2011, 18(3): 646-656
[13] Xu Y, Wang L, Zhou G, et al. An effective shuffled frog leaping algorithm for solving hybrid flow-shop scheduling problem[C]// Proceedings of the 9th International Conference on Advanced Intelligent Computing. Berlin: Springer Verlag, 2013: 560-567
[14] Dong W S, Yao X. Unified eigen analysis on multivariate Gaussian based estimation of distribution algorithms [J]. Information Sciences, 2011, 178(15): 3000-3023
[15] Muhlenbein H. The equation for response to selection and its use for prediction [J]. Evolutionary Computation, 2012, 5(3): 303-346
[16] Kastegar R. On the optimal convergence probability of univariate estimation of distribution algorithms [J]. Evolutionary Computation, 2013, 19(2): 225-248
[17] The Cloud Lab. Cloudsim [EB/OL]. [2012-08-15]. http://www.cloudbus.org/cloudsim