可调节模糊粗糙集:模型与属性约简

宋晶晶[1] 杨习贝[1,2,3] 戚 湧[4] 祁云嵩[1]
(江苏科技大学计算机科学与工程学院 镇江 212003)[1] (人工智能四川省重点实验室 自贡 643000)[2]
(高维信息智能感知与系统教育部重点实验室 南京 210094)[3]
(南京理工大学经济管理学院 南京 210094)[4]

摘 要 模糊粗糙集是经典粗糙集为适应实际应用需求所进行的拓展,然而目前很多的模糊粗糙集模型都仅仅使用多个二元关系的简单融合方式,不具备调节功能。为解决这一问题,使用参数化的二元算子,提出了一种可调节的模糊粗糙集模型。在此基础上,将近似质量作为度量标准,使用启发式算法来求解可调节模糊粗糙集的约简。最后对可调节模糊粗糙集的近似质量和约简与强模糊粗糙集、弱模糊粗糙集的结果进行了比较分析。实验结果表明,可调节模糊粗糙集通过使用不同的参数,具有很好的调节作用,是强模糊粗糙集和弱模糊粗糙集的一种泛化形式。

关键词 近似质量,决策系统,模糊粗糙集,约简
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.040

Adjustable Fuzzy Rough Set:Model and Attribute Reduction

SONG Jing-jing[1] YANG Xi-bei[1,2,3] QI Yong[4] QI Yun-song[1]
(School of Computer Science and Engineering,Jiangsu University of Science and Technology,Zhenjiang 212003,China)[1]
(Artificial Intelligence Key Laboratory of Sichuan Province,Zigong 643000,China)[2]
(Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information,Ministry of Education,Nanjing 210094,China)[3]
(School of Economics and Management,Nanjing University of Science and Technology,Nanjing 210094,China)[4]

Abstract Fuzzy rough set is an extension of classical rough set by considering requirements of the practical applications. However,many existing fuzzy rough set models only use simple fusions of a set of binary relations,and these fusions are not adjustable. To solve such problem,an adjustable fuzzy rough set was proposed by using a parameterized binary operator. Moreover,the approximate quality was regarded as a measurement and then the heuristic algorithm was used to calculate the reduction of adjustable fuzzy rough set. Finally,the approximate quality and the reduction of adjustable fuzzy rough set were compared with those of the strong fuzzy rough set and the weak fuzzy rough set respectively. The experimental results show that adjustable fuzzy rough set is a generalization of both strong and weak fuzzy rough sets.

Keywords Approximation quality,Decision system,Fuzzy rough set,Reduction

1 引言

经典粗糙集是由波兰学者Pawlak[1]于1982年提出的一种刻画不确定问题的数学工具。由于经典的粗糙集只能处理离散型数据,而实际的数据集中存在很多复杂类型的数据,为此Dubois和Prade[2]在1990年将经典粗糙集中的等价关系推广至模糊等价关系(满足自反性、对称性和max-min传递性的模糊关系),使用一对min和max算子对下近似和上近似进行定义,从而为模糊粗糙集的发展奠定了基石。

近年来,国内外学者对模糊粗糙集理论的研究主要集中在3个方面:模糊粗糙集的公理化方法[3-8]、模糊粗糙集的构造性方法[9-18]以及模糊粗糙集在模式识别、数据挖掘、特征选择等各领域的应用[19-26]。例如,在Dubois与Prade的模糊粗糙集的基础上,Morsi和Yakout[3]引入模糊T-相似关系对模糊粗糙集进行了重新定义;由于模糊粗糙集中的min和max算子是一对特殊的t-模和t-余模[27],Yeung等[28]引入更一般化的t-模和t-余模S对模糊粗糙集进行了重新定义;Hu等[9,10]将高斯核函数用于计算对象之间的模糊等价关系,提出了高斯核模糊粗糙集;Chen等[13]从几何角度,将模糊粗糙集的隶属度解释为Krein空间。作为粗糙集理论的重要扩

到稿日期:2014-01-28 返修日期:2014-05-04 本文受国家自然科学基金(61100116,61272419),江苏省自然科学基金(BK2011492,BK2012700,BK20130471),高维信息智能感知与系统教育部重点实验室(南京理工大学)开放基金(30920130122005),人工智能四川省重点实验室开放基金重点课题(2013RYJ03),江苏省高校自然科学基金(13KJB520003,13KJD520008)资助。

宋晶晶(1990—),女,硕士生,主要研究方向为粗糙集理论;杨习贝(1980—),男,博士后,副教授,硕士生导师,主要研究方向为粒计算、粗糙集、数据挖掘与知识发现,E-mail:zhenjiangyangxibei@163.com(通信作者);戚 湧(1970—),男,博士,教授,博士生导师,主要研究方向为人工智能与信息处理;祁云嵩(1964—),男,博士,教授,硕士生导师,主要研究方向为高维信息处理。

展，模糊粗糙集已俨然受到了众多学者的高度重视。

值得注意的是，从二元关系的角度来看，诸多学者对于模糊粗糙集的刻画是没有调节能力的。其中，Hu 等人[9,10]提出的高斯核模糊粗糙集中的参数虽然具有调节作用，但不能针对多个二元关系进行调节，以根据实际应用的多样需求，做出变化。所以，研究一种可根据实际工程需要进行适当调节的模糊粗糙集方法显得十分必要。庆幸的是在 1975 年，Fung 和 Fu[29]提出了一种可调节算子，通过给定的一个参数，可以在 min 和 max 算子之间进行调节。由于可调节算子可视作是 min 和 max 算子的一种扩展形式，而 min 和 max 算子则是可调节算子的特例，因此可以通过可调节算子来实现模糊粗糙集的调节能力。

本文将 Fung 和 Fu 的参数化算子引入到二元模糊关系中，使得模糊粗糙近似具备可调节能力，提出了一种新的可调节的模糊粗糙集模型。进一步地，利用文献[9]中的高斯核函数求得模糊关系，并使用近似质量作为度量，采用启发式算法求解可调节模糊粗糙集的约简，最后对可调节模糊粗糙集与强、弱模糊粗糙集进行了分析。

2 基本概念

2.1 模糊信息粒化

一般地，一个三元组 $DS=(U,AT\cup D,V)$ 被称为一个决策系统，其中非空有限集合 U 是所有研究对象的合集，称为论域；非空有限集合 AT 是所有条件属性的合集；非空有限集合 D 是所有决策属性的合集且 $AT\cap D=\varnothing$；V 是所有属性的值域，即 $V=\bigcup\limits_{a\in AT\cup D}V_a$，$\forall x\in U$，$a(x)$ 表示对象 x 在属性 a 上的取值。

令 $U\neq\varnothing$ 为一论域，F 是从 U 到区间值 $[0,1]$ 的映射，即 $F:U\to[0,1]$，称 F 为论域 U 上的模糊集[2]，$F(x)\in[0,1]$ 是 x 隶属于模糊集 F 的程度。其中，当 F 是经典集合时，其隶属度取值为 1 或 0。便于讨论，本文将论域 U 上所有模糊集的集合表示为 $\boldsymbol{F}(U)$，模糊子集 $R\in\boldsymbol{F}(U\times U)$ 是一个模糊关系，称 (U,R) 为一个模糊近似空间。

论域 U 上的模糊关系 R 是线性的当且仅当 $\forall x\in U,\exists y\in U$ 使得 $R(x,y)=1$；模糊关系 R 是自反的当且仅当 $R(x,x)=1(\forall x\in U)$；模糊关系 R 是对称的当且仅当 $R(x,y)=R(y,x)(\forall x,y\in U)$；模糊关系 R 是传递的当且仅当 $R(x,y)\wedge R(y,z)\leqslant R(x,z)(\forall x,y,z\in U)$。文中所讨论的模糊关系至少是自反的。

给定一个决策系统 $DS=(U,AT\cup D,V)$，$\forall a_i\in AT$，根据属性 a_i 所诱导的模糊关系可以用一个模糊关系矩阵 R^i 进行表示：

$$R^i=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

其中，n 表示论域 U 中对象的个数，$r_{ij}\in[0,1]$ 表示对象 x_i 与 x_j 之间的相似度（$\forall x_i,x_j\in U$）。

针对模糊关系，可定义如下所示的运算：

1）$R^1=R^2\Leftrightarrow R^1(x,y)=R^2(x,y),\forall x,y\in U$；

2）$R=R^1\cup R^2\Leftrightarrow R=\max\{R^1(x,y),R^2(x,y),\forall x,y\in U\}$；

3）$R=R^1\cap R^2\Leftrightarrow R=\min\{R^1(x,y),R^2(x,y),\forall x,y\in U\}$；

4）$R^1\subseteq R^2\Leftrightarrow R^1(x,y)\leqslant R^2(x,y),\forall x,y\in U$。

以粒计算的观点来看，论域 U 上的一个模糊关系 R 可以诱导出一个模糊信息粒化，且该模糊信息粒化是一个邻域系统，表示为：

$$N(R)=\{G_R(x_1),G_R(x_2),\cdots,G_R(x_n)\}(x_1,x_2,\cdots,x_n\in U)$$

其中，$G_R(x_i)=r_{i1}/x_1+r_{i2}/x_2+\cdots+r_{in}/x_n$ 表示由对象 x_i 诱导的模糊信息粒（也可以视作对象 x_i 的邻域）[30]。

2.2 模糊粗糙集

定义 1 令 U 为论域，R 是论域 U 上的一个模糊关系，$\forall F\in\boldsymbol{F}(U)$，$F$ 在模糊近似空间 (U,R) 中的下近似 $\underline{R}(F)$ 和上近似 $\overline{R}(F)$ 是 U 上的一对模糊集，$\forall x\in U$，其隶属度函数分别为：

$$\underline{R}(F)(x)=\bigwedge_{y\in U}((1-R(x,y))\vee F(y)) \tag{1}$$

$$\overline{R}(F)(x)=\bigvee_{y\in U}(R(x,y)\wedge F(y)) \tag{2}$$

称 $(\underline{R}(F),\overline{R}(F))$ 为 F 的一个模糊粗糙集[2]。

在决策系统 $DS=(U,AT\cup D,V)$ 中，假定论域 U 中条件属性集 AT 的个数为 m，$\forall a_i\in AT(1\leqslant i\leqslant m)$，在论域 U 上都可以构建一个模糊关系 R^i，所以针对整个决策系统来说，可考虑如下所示的两种特殊模糊关系：

1）模糊关系 $R^S=R^1\cap R^2\cap\cdots\cap R^m$，称 R^S 为强模糊关系，用模糊关系矩阵 $M(R^S)$ 表示。此时，$\forall x_i,x_j\in U$，对象 x_i 与 x_j 之间的相似度最小，为方便讨论，用 r_{ij}^s 表示。

2）模糊关系 $R^W=R^1\cup R^2\cup\cdots\cup R^m$，称 R^W 为弱模糊关系，用模糊关系矩阵 $M(R^W)$ 表示。此时，$\forall x_i,x_j\in U$，对象 x_i 与 x_j 之间的相似度最大，为方便讨论，用 r_{ij}^w 表示。

定义 2 令 U 为论域，R^S 是论域 U 上的强模糊关系，$\forall F\in\boldsymbol{F}(U)$，$F$ 在模糊近似空间 (U,R^S) 中的下近似 $\underline{R^S}(F)$ 和上近似 $\overline{R^S}(F)$ 是 U 上的一对模糊集，$\forall x\in U$，其隶属度函数分别为：

$$\underline{R^S}(F)(x)=\bigwedge_{y\in U}((1-R^S(x,y))\vee F(y)) \tag{3}$$

$$\overline{R^S}(F)(x)=\bigvee_{y\in U}(R^S(x,y)\wedge F(y)) \tag{4}$$

称 $(\underline{R^S}(F),\overline{R^S}(F))$ 为 F 的一个强模糊粗糙集。

定义 3 令 U 为论域，R^W 是论域 U 上的弱模糊关系，$\forall F\in\boldsymbol{F}(U)$，$F$ 在模糊近似空间 (U,R^W) 中的下近似 $\underline{R^W}(F)$ 和上近似 $\overline{R^W}(F)$ 是 U 上的一对模糊集，$\forall x\in U$，其隶属度函数分别为：

$$\underline{R^W}(F)(x)=\bigwedge_{y\in U}((1-R^W(x,y))\vee F(y)) \tag{5}$$

$$\overline{R^W}(F)(x)=\bigvee_{y\in U}(R^W(x,y)\wedge F(y)) \tag{6}$$

称 $(\underline{R^W}(F),\overline{R^W}(F))$ 为 F 的一个弱模糊粗糙集。

3 可调节模糊粗糙集

3.1 可调节模糊粗糙集模型

不失一般性，假设 m 是一个指标集，$\forall i\in m$，$\forall r_i\in[0,1]$，一个参数化算子定义如下[29]：

$$\bigvee_{i\in m}^{\lambda}r_i=\begin{cases}\bigvee\limits_{i\in m}r_i, & \{r_i:i\in m\}\subseteq[0,\lambda]\\ \bigwedge\limits_{i\in m}r_i, & \{r_i:i\in m\}\subseteq[\lambda,1]\\ \lambda, & \text{其他}\end{cases} \tag{7}$$

其中 $\lambda\in[0,1]$。

特别地，当指标集 $m=2$ 时，$\forall r_1,r_2\in[0,1]$，有

$$r_1\overset{\lambda}{\vee}r_2=\begin{cases}r_1\vee r_2, & r_1,r_2\in[0,\lambda]\\ r_1\wedge r_2, & r_1,r_2\in[\lambda,1]\\ \lambda, & \text{其他}\end{cases}\tag{8}$$

显然，参数化算子 $\overset{\lambda}{\underset{i\in m}{\vee}}$ 是最大算子($\vee$)和最小算子($\wedge$)的一种广义化形式。可以通过设置 λ 的值来对 r_1 与 r_2 的运算结果进行调节。特别地，当 $\lambda=0$ 时，有 $\overset{0}{\underset{i\in m}{\vee}}=\wedge$；当 $\lambda=1$ 时，有 $\overset{1}{\underset{i\in m}{\vee}}=\vee$。

命题 1 令 m 是一个指标集，$\forall i\in m,\forall r_i\in[0,1]$，有 $\underset{i\in m}{\wedge}r_i\leqslant\overset{\lambda}{\underset{i\in m}{\vee}}r_i\leqslant\underset{i\in m}{\vee}r_i$。

证明：根据式(7)，可做如下 3 种情况的讨论：

1)当 $\{r_i:i\in m\}\subseteq[0,\lambda]$ 时，有 $\underset{i\in m}{\wedge}r_i\leqslant\overset{\lambda}{\underset{i\in m}{\vee}}r_i=\underset{i\in m}{\vee}r_i$。

2)当 $\{r_i:i\in m\}\subseteq[\lambda,1]$ 时，有 $\underset{i\in m}{\wedge}r_i=\overset{\lambda}{\underset{i\in m}{\vee}}r_i\leqslant\underset{i\in m}{\vee}r_i$。

3)否则，显然 $\underset{i\in m}{\wedge}r_i\leqslant\overset{\lambda}{\underset{i\in m}{\vee}}r_i=\lambda\leqslant\underset{i\in m}{\vee}r_i$ 成立。

在决策系统 $DS=(U,AT\cup D,V)$ 中，若条件属性的个数为 m，$\forall a_i\in AT(1\leqslant i\leqslant m)$，在论域 U 上都可以构建一个模糊关系 R^i。$\forall x_i,x_j\in U$ 根据参数化算子的定义，将 m 个模糊关系 R^i 中的对象 x_i 与 x_j 之间的相似度用参数 λ 进行调节，调整之后的相似度用 r_{ij}^{λ} 表示，则有 $r_{ij}^{\lambda}=\overset{\lambda}{\underset{i\in m}{\vee}}r_{ij}$，所以将由 r_{ij}^{λ} 构成的模糊关系矩阵称为可调节的模糊关系矩阵并记为 $M(R^{\lambda})$，将模糊关系矩阵 $M(R^{\lambda})$ 表示的模糊关系 R^{λ} 称为可调节模糊关系。

命题 2 令 $DS=(U,AT\cup D,V)$ 为一个决策系统，可调节模糊关系 R^{λ} 可以由强模糊关系 R^S 与弱模糊关系 R^W 通过参数 λ 进行调节得到，即 $\forall x_i,x_j\in U$，有 $r_{ij}^{\lambda}=r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}$。

证明：因为 $r_{ij}^{s}\leqslant r_{ij}^{w}$，所以可做如下 3 种情况的讨论：

1)$\forall x_i,x_j\in U$，当 $\{r_{ij}^{s},r_{ij}^{w}\}\subseteq[0,\lambda]$ 时，根据式(7)，有 $r_{ij}^{\lambda}=\overset{\lambda}{\underset{i\in m}{\vee}}r_{ij}=r_{ij}^{w}$，再根据式(8)，有 $r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}=r_{ij}^{w}$，故 $r_{ij}^{\lambda}=r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}$。

2)$\forall x_i,x_j\in U$，当 $\{r_{ij}^{s},r_{ij}^{w}\}\subseteq[\lambda,1]$ 时，根据式(7)，有 $r_{ij}^{\lambda}=\overset{\lambda}{\underset{i\in m}{\vee}}r_{ij}=r_{ij}^{s}$，再根据式(8)，有 $r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}=r_{ij}^{s}$，故 $r_{ij}^{\lambda}=r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}$。

3)否则，根据式(7)，有 $r_{ij}^{\lambda}=\overset{\lambda}{\underset{i\in m}{\vee}}r_{ij}=\lambda$，再根据式(8)，有 $r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}=\lambda$，故 $r_{ij}^{\lambda}=r_{ij}^{s}\overset{\lambda}{\vee}r_{ij}^{w}$。

定义 4 令 U 为论域，R^{λ} 是 U 上的一个可调节模糊关系，$\forall F\in\boldsymbol{F}(U)$，$F$ 在模糊近似空间 (U,R^{λ}) 中的下近似 $\underline{R^{\lambda}}(F)$ 和上近似 $\overline{R^{\lambda}}(F)$ 是 U 上的一对模糊集，$\forall x\in U$，其隶属度函数分别为：

$$\underline{R^{\lambda}}(F)(x)=\underset{y\in U}{\wedge}((1-R^{\lambda}(x,y))\vee F(y))\tag{9}$$

$$\overline{R^{\lambda}}(F)(x)=\underset{y\in U}{\vee}(R^{\lambda}(x,y)\wedge F(y))\tag{10}$$

称 $(\underline{R^{\lambda}}(F),\overline{R^{\lambda}}(F))$ 为 F 的一个可调节模糊粗糙集。

定理 1 令 $DS=(U,AT\cup D,V)$ 为一个决策系统，$\forall F\in\boldsymbol{F}(U)$，有

$$\underline{R^0}(F)(x)=\underline{R^S}(F)(x),\overline{R^0}(F)(x)=\overline{R^S}(F)(x)\tag{11}$$

$$\underline{R^1}(F)(x)=\underline{R^W}(F)(x),\overline{R^1}(F)(x)=\overline{R^W}(F)(x)\tag{12}$$

证明：根据式(7)，可知 $\overset{0}{\underset{i\in m}{\vee}}=\wedge$，再根据定义 4 可知，$\forall x\in U$，$\underline{R^0}(F)(x)=\underset{y\in U}{\wedge}((1-R^0(x,y))\vee F(y))=\underset{y\in U}{\wedge}((1-R^S(x,y))\vee F(y))=\underline{R^S}(F)(x)$。

根据式(7)，可知 $\overset{0}{\underset{i\in m}{\vee}}=\wedge$，再根据定义 4 可知，$\forall x\in U$，$\overline{R^0}(F)(x)=\underset{y\in U}{\vee}(R^0(x,y)\wedge F(y))=\underset{y\in U}{\vee}(R^S(x,y)\wedge(F(y))=\overline{R^S}(F)(x)$。

同理，不难证得 $\underline{R^1}(F)(x)=\underline{R^W}(F)(x)$ 与 $\overline{R^1}(F)(x)=\overline{R^W}(F)(x)$ 成立。

由定理 1 可知，强模糊粗糙集和弱模糊粗糙集是可调节模糊粗糙集的两种特殊情形。虽然在定义 4 中，参数的引入引起了用户主观性的增加，但同时由定理 1 可以看出，我们也得到了一种广义化的模糊粗糙集表现形式。

定理 2 令 $DS=(U,AT\cup D,V)$ 是一个决策系统，$\forall F\in\boldsymbol{F}(U)$，有

$$\underline{R^W}(F)\subseteq\underline{R^{\lambda}}(F)\subseteq\underline{R^S}(F)\tag{13}$$

$$\overline{R^S}(F)\subseteq\overline{R^{\lambda}}(F)\subseteq\overline{R^W}(F)\tag{14}$$

证明：根据命题 1，定理 2 显然成立。

由定理 2 可以看出，可调节模糊粗糙集的下、上近似集分别介于强模糊粗糙集和弱模糊粗糙集的下、上近似集之间。

3.2 近似质量

近似质量是利用模糊粗糙集来评估 $U/IND(D)$ 中的确定性程度。在本小节中，我们将近似质量引入到强模糊粗糙集、弱模糊粗糙集和可调节模糊粗糙集中，并对其关系进行讨论。

定义 5 令 $DS=(U,AT\cup D,V)$ 为一个决策系统，由决策属性 D 诱导的论域上的划分为 $U/IND(D)=\{d_1,d_2,\cdots,d_l\}$，则模糊粗糙集的近似质量定义如下：

$$\gamma(AT,D)=\frac{|\bigcup_{i=1}^{l}\underline{Rd_i}|}{|U|}\tag{15}$$

类似于定义 5，强模糊粗糙集、弱模糊粗糙集和可调节模糊粗糙集的近似质量分别定义如下：

$$\gamma^s(AT,D)=\frac{|\bigcup_{i=1}^{l}\underline{R^Sd_i}|}{|U|}\tag{16}$$

$$\gamma^w(AT,D)=\frac{|\bigcup_{i=1}^{l}\underline{R^Wd_i}|}{|U|}\tag{17}$$

$$\gamma^{\lambda}(AT,D)=\frac{|\bigcup_{i=1}^{l}\underline{R^{\lambda}d_i}|}{|U|}\tag{18}$$

定理 3 令 $DS=(U,AT\cup D,V)$ 为一个决策系统，$\forall F\in\boldsymbol{F}(U)$，有

$$\gamma^w(AT,D)\leqslant\gamma^{\lambda}(AT,D)\leqslant\gamma^s(AT,D)\tag{19}$$

证明：根据定理 2，定理 3 显然成立。

通过定理 3 可以看出，可调节模糊粗糙集的近似质量介于强模糊粗糙集和弱模糊粗糙集的近似质量之间。

3.3 属性约简

属性约简是粗糙集的核心内容之一，它根据特定需要，将数据集中的冗余属性删除。在此，我们基于近似质量对可调节模糊粗糙集的属性约简问题进行研究。

定义 6 令 $DS=(U,AT\cup D,V)$ 为一个决策系统，$\gamma\in$

$[0,1]$，$\forall A \subseteq AT$，A 是 DS 的一个近似质量约简当且仅当 $\gamma^{\lambda}(A,D)=\gamma^{\lambda}(AT,D)$ 并且 $\forall B \subset A$，都有 $\gamma^{\lambda}(B,D) \neq \gamma^{\lambda}(AT,D)$。

由定义 6 可以看出，决策系统 DS 中的一个近似质量约简是保持可调节模糊粗糙集的近似质量不变的最小属性子集。

令 $DS=(U,AT \cup D,V)$ 是一个决策系统，$\lambda=[0,1]$，$\forall A \subseteq AT$，$\forall a_i \in A$ 属性 a_i 的重要度定义如下：

$$Sig_{in}(a_i,A,D)=abs(\gamma^{\lambda}(A,D)-\gamma^{\lambda}(A-a_i,D))$$

其中，abs 代表的是绝对值。$Sig_{in}(a_i,A,D)$ 用来反映从当前条件属性集 A 中删除属性 a_i 后的近似质量的变化。相应地，定义

$$Sig_{out}(a_i,A,D)=abs(\gamma^{\lambda}(A \cup a_i,D)-\gamma^{\lambda}(A,D))$$

$Sig_{out}(a_i,A,D)$ 用来反映从当前条件属性集 A 中增加属性 a_i 后的近似质量的变化。

根据上述属性重要度的定义，不难给出如下所示的启发式算法，用于求解可调节模糊粗糙集的一个近似质量约简。

算法 启发式算法

输入：决策系统 DS，λ；

输出：可调节模糊粗糙集的一个近似质量约简 red.

1. 计算 $\gamma^{\lambda}(AT,D)$；
2. red ←∅；
3. $\forall a_i \in AT$，计算属性 a_i 的重要度 $Sig_{in}(a_i,AT,D)$；
4. 若 $Sig(a_j,AT,D)=\max\{Sig(a_i,AT,D): \forall a_i \in AT\}$，则 red←$a_j$，计算 $\gamma^{\lambda}(red,D)$；
5. 若 $\gamma^{\lambda}(red,D) \neq \gamma^{\lambda}(AT,D)$，则重复以下循环，否则转入步骤 6；
   (1) $\forall a_i \in AT-red$，计算 $Sig_{out}(a_i,red,D)$；
   (2) 若 $Sig_{out}(a_j,red,D)=\max\{Sig(a_i,red,D): \forall a_i \in AT-red\}$，则 red←$a_j$；
   (3) 计算 $\gamma^{\lambda}(red,D)$.
6. $\forall a_i \in red$，若 $\gamma^{\lambda}(red-a_i,D)=\gamma^{\lambda}(red,D)$，则 red=red−$a_i$；
7. 输出 red.

4 实验分析

本节首先通过实验来分析强模糊粗糙集、弱模糊粗糙集的近似质量以及随着参数 λ 的不同，可调节模糊粗糙集的近似质量的变化。为此，我们从 UCI 公共数据集上下载了 10 组数据集，数据集的基本信息如表 1 所列。

表 1 实验数据基本信息

序号	数据集名称	对象数	特征数	决策类数
1	Connectionist Bench(Sonar, Mines vs. Rocks)	208	60	2
2	Ecoli	336	7	8
3	Glass Identification	214	10	6
4	Ionosphere	351	34	2
5	Iris	150	4	3
6	Libras Movement	360	90	4
7	Pima Indians Diabetes	768	8	2
8	Seed	210	7	3
9	Statlog(Heart)	270	13	2
10	Wine	178	13	3

表 2 3 种模糊粗糙集的近似质量

序号	强模糊粗糙集	不同 λ 取值的可调节模糊粗糙集										弱模糊粗糙集
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
1	1.0000	1.0000	0.9000	0.8000	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000	0	0
2	0.8669	0.8669	0.8183	0.7475	0.6673	0.5813	0.4901	0.3960	0.2994	0.2000	0.1000	0	0
3	0.9638	0.9638	0.8850	0.7950	0.6983	0.5998	0.5000	0.4000	0.3000	0.2000	0.1000	0	0
4	0.9984	0.9984	0.9000	0.8000	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000	0	0
5	0.9587	0.9587	0.8795	0.7925	0.6987	0.6000	0.5000	0.4000	0.3000	0.2000	0.1003	0.0010	0.0010
6	0.9863	0.9863	0.8877	0.7890	0.6904	0.5918	0.4932	0.3945	0.2959	0.1973	0.0986	0	0
7	0.9055	0.9055	0.8566	0.7783	0.6899	0.5955	0.4988	0.3994	0.2997	0.2000	0.1000	0	0
8	0.9689	0.9689	0.8842	0.7904	0.6941	0.5965	0.4987	0.3997	0.3000	0.2000	0.1000	0.0002	0.0002
9	0.9994	0.9994	0.9000	0.8000	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000	0	0
10	1.0000	1.0000	0.9000	0.8000	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000	0	0

表 3 3 种模糊粗糙集的属性约简

序号	原特征数	约简后的特征数											
		强模糊粗糙集约简后的特征数	不同 λ 取值的可调节模糊粗糙集										弱模糊粗糙集约简后的特征数
			0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
1	60	48	48	6	6	5	5	5	5	4	4	3	10	10
2	7	7	7	6	6	6	6	6	6	6	4	4	1	1
3	10	10	10	9	8	7	7	6	5	4	4	4	3	3
4	34	32	32	8	7	6	6	6	6	6	6	5	1	1
5	4	4	4	4	4	4	4	4	3	3	3	3	2	2
6	90	69	69	9	8	8	6	6	5	3	4	4	4	4
7	8	8	8	8	8	8	8	8	8	6	6	6	2	2
8	7	7	7	6	6	6	6	5	4	4	3	4	7	7
9	13	13	13	9	8	8	7	6	6	6	5	5	1	1
10	13	13	13	6	5	5	4	4	4	4	4	3	4	4

由第 2 节可知，根据数据集中的每一个属性，都可以构建一个模糊关系。本文使用高斯核函数来计算数据集中每一个属性下对象 x_i 与 x_j（$\forall x_i,x_j \in U$）之间的相似度 r_{ij}，即 $r_{ij}=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$[9]。其中，核函数中的参数 σ 的选择非常重要。σ 越大，所诱导的模糊信息粒度越大。当 σ 趋向于无穷时，$\forall x_i,x_j \in U$，对象之间的相似度趋向于 1，对象之间变得不可区分。在本节的实验中，均将参数 σ 设置为 0.05[9]。

在研究强、弱和可调节这 3 种模糊粗糙集的近似质量实验中，将可调节参数 λ 的值设置为从 0 到 1 每次递增 0.1，进行实验。实验结果如表 2 所列。在实际的应用中，可以根据用户的需要来设置参数 λ。

从表 2 可以看出，当可调节参数 $\lambda=0$ 时，可调节模糊粗糙集的近似质量与强模糊粗糙集的近似质量相等；当可调节参数 $\lambda=1$ 时，可调节模糊粗糙集的近似质量与弱模糊粗糙集的近似质量相等。当 $0<\lambda<1$ 时，可调节模糊粗糙集的近似质量在强、弱模糊粗糙集的近似质量之间。该现象验证了定理 3 的理论结果。

我们仍然采用表 1 中的 10 个数据集，根据启发式算法，求可调节模糊粗糙集的近似质量约简，并与强、弱模糊粗糙集的近似质量约简进行对比，实验结果如表 3 所列。

从表 3 中可以看出，当可调节参数 $\lambda=0$ 时，可调节模糊粗糙集的约简属性个数与强模糊粗糙集的约简属性个数是相等的；当可调节参数 $\lambda=1$ 时，可调节模糊粗糙集的约简属性个数与弱模糊粗糙集的约简属性个数是相等的。当 $0<\lambda<1$ 时，可调节模糊粗糙集的约简属性个数介于强模糊粗糙集与弱模糊粗糙集的约简属性个数之间。

通过上述两组实验，可以得出以下结论：

1)随着可调节参数 λ 值的不断增大，可调节模糊粗糙集的近似质量单调递减。

2)可调节模糊粗糙集的近似质量约简中的属性个数随着可调节参数 λ 的不断增大，大致呈现出单调递减的趋势。

3)当可调节参数 λ 较小时，近似质量较大，要保持近似质量不变，则所需得到的属性个数就比较多，因此所得到近似质量约简的属性个数也比较多。

结束语 本文针对如何使模糊粗糙集具有调节性的问题进行了讨论，通过引入二元可调节算子，提出了可调节模糊粗糙集。然后讨论了可调节模糊粗糙集的近似质量与强、弱模糊粗糙集之间的关系。最后，通过实验分析，研究了可调节模糊粗糙集的约简结果与可调节参数之间的关系，并且讨论了可调节模糊粗糙集的约简与强、弱模糊粗糙集约简之间的关系。

在下一步的工作中，笔者将从以下两个方面进行研究：

1)结合机器学习中的模型选择、超参数学习等相关内容，对参数 σ 的设置进行深入研究，以便寻找出更加适合实际应用的 σ 值。

2)使用多种约简算法对可调节模糊粗糙集进行属性约简的对比分析，讨论它们的优劣，从而可以根据不同需要，选择不同的约简算法，进行问题求解。

参 考 文 献

[1] Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data[M]. Kluwer Academic Publishers, Dordrecht, Boston, London, 1991

[2] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17: 191-209

[3] Morsi N N, Yakout M M. Axiomatics for fuzzy rough sets[J]. Fuzzy Sets and Systems, 1998, 100: 327-342

[4] Mi J S, Leung Y, Zhao H Y, et al. Generalized fuzzy rough sets determined by a triangular norm[J]. Information Sciences, 2008, 178(16): 3203-3213

[5] She Y H, Wang G Y. An axiomatic approach of fuzzy rough sets based on residuated lattices[J]. Computers & Mathematics with Applications, 2009, 58(1): 189-201

[6] Chen D G, Yang Y Y, Wang H. Granular computing based on fuzzy similarity relations[J]. Soft Computing, 2011, 15(6): 1161-1172

[7] Wang C Y, Hu B Q. Fuzzy rough sets based on generalized residuated lattices[J]. Information Sciences, 2013, 248: 31-49

[8] Liu G L. Using one axiom to characterize rough set and fuzzy rough set approximations[J]. Information Sciences, 2013, 223: 285-296

[9] Hu Q H, Zhang L, Chen D G, et al. Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications[J]. International Journal of Approximate Reasoning, 2010, 51(4): 453-471

[10] Hu Q H, Yu D R, Pedrycz W, et al. Kernelized fuzzy rough sets and their applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(11): 1649-1667

[11] He Q, Wu C X, Chen D G, et al. Fuzzy rough set based attribute reduction for information systems with fuzzy decisions[J]. Knowledge-Based Systems, 2011, 24(5): 689-696

[12] Huang B, Li H X, Wei D K. Dominance-based rough set model in intuitionistic fuzzy information systems[J]. Knowledge-Based Systems, 2012, 28: 115-123

[13] Chen D G, Kwong S, He Q, et al. Geometrical interpretation and applications of membership functions with fuzzy rough sets[J]. Fuzzy Sets and Systems, 2012(193): 122-135

[14] Wu W Z, Leung Y, Shao M W. Generalized fuzzy rough approximation operators determined by fuzzy implicators[J]. International Journal of Approximate Reasoning, 2013, 54(9): 1388-1409

[15] Huang B, Zhang Y L, Li H X, et al. A dominance intuitionistic fuzzy-rough set approach and its applications[J]. Applied Mathematical Modelling, 2013, 37(12/13): 7128-7141

[16] Wang C Y, Hu B Q. Fuzzy rough sets based on generalized residuated lattices[J]. Information Sciences, 2013, 248: 31-49

[17] Dai J H, Tian H W, Fuzzy rough set model for set-valued data [J]. Fuzzy Sets and Systems, 2013, 229: 54-68

[18] Yao Y Q, Mi J S, Li Z J. A novel variable precision (θ,σ)-fuzzy rough set model based on fuzzy granules[J]. Fuzzy Sets and Systems, 2014, 236: 58-72

[19] Chen D G, Zhao S Y. Local reduction of decision system with fuzzy rough sets[J]. Fuzzy Sets and Systems, 2010, 161(13): 1871-1883

[20] Hu Q H, Shang A, Yu D R. Soft fuzzy rough sets for robust feature evaluation and selection[J]. Information Sciences, 2010, 180(22): 4383-4400

[21] Maji P. Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data[J]. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 2011, 41(1): 222-233

[22] Chen D G, Hu Q H, Yang Y P. Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets[J]. Information Sciences, 2011, 181(23): 5169-5179

[23] Chen D G, Zhang L, Zhao S Y, et al. A novel algorithm for finding reducts with fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(2): 385-389

[24] Dai J H, Xu Q. Attribute selection based on information gain ra-

tio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing,2013,13(1):211-221
[25] 徐菲菲,魏莱,杜海洲,等. 一种基于互信息的模糊粗糙集分类特征基因快速选取方法[J]. 计算机科学,2013,40(7):216-235
[26] 曾安平,李天瑞,罗川. 高斯核模糊粗糙集中对象集变化时近似集增量更新方法研究[J]. 计算机科学,2013,40(7):173-177
[27] Klement E P, Mesiar R, Pap E. Triangular norms[M]. Kluwer Academic Publishers,2001
[28] Yeung D S,Chen D G,Tsang E C C,et al. On the generalization of fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2005,13(3):343-361
[29] Fung L W, Fu K S. An axiomatic approach to relational decision-making in a fuzzy environment[C]// Zadeh L A,Fu K S,Tanaka K,et al. , eds. Fuzzy Sets and Decision Processes. New York: Academic Press,1975:227-256
[30] Qian Y H,Wu W Z,Dang C Y. Information granularity in fuzzy binary GrC model[J]. IEEE Transactions on Fuzzy Systems, 2011,19(2):253-264

(上接第 171 页)

询时的缩略用法,这样的词进行同义扩展的空间较大,因此性能提升较大。共现词扩展则正好相反,与高频查询词共现的词与查询词具有很强的关联性,扩展查询效果较好;而与低频查询词共现的词与查询词不具备强关联性,利用其进行扩展反而引起语义的迁移。由于隐马尔可夫模型在进行用户意图预测的过程中,结合了同义词扩展和共现词扩展的优点,因此在两种环境中查询的性能均有一定提升。同时,利用隐马尔可夫模型扩展也具有共现词扩展的不足,即在训练语料量较少的情况下,性能提升有限,部分查询扩展查询后准确率甚至不如未扩展的准确率。

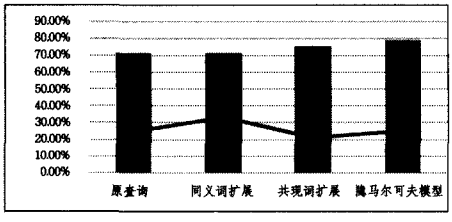

图 4 扩展查询准确率

需要特别指出的是,利用隐马尔可夫模型进行扩展查询的优势在于其意图预测。例如在我们的实验中,用户输入的"清华"被扩展成了"清华大学 官网"。这一扩展查询代表了大多数用户的查询意图,其返回的清华大学官方网站链接排在搜索引擎返回页面的首位,而其他的返回页面由于受到"官网"这个扩展词的限制,相关度急剧下降,这使得我们的实验结果在一定程度上有所损失,但并不影响将最相关的页面返回给用户的初衷。

结束语 扩展查询是提高搜索引擎性能的重要手段之一。本文给出了一种利用隐马尔可夫模型进行扩展查询的方法,该方法综合了基于同义词表扩展查询和基于共现词表扩展查询的优点,利用大规模用户查询日志进行模型训练,实现了对用户查询中潜在意图的预测。由该方法得到的扩展查询模型在中高频查询词中取得了较好的效果,而对于低频查询词则仍需进一步改进。

由于用户输入的查询词数较少,因此本文提出的基于最长公共子串的对齐方式系统开销较少,这让模型在大规模查询日志上进行训练成为可能。

参 考 文 献

[1] Crouch C J. A cluster-based approach to thesaurus construction [C]// Eleventh International ACMSIGIR Conference on Research and Development in Information Retrieval. 1988:309-320
[2] Blondel V D,Senellart P P. Automatic extraction of synonyms in a dictionary [R]. Presented at the Text Mining Workshop,2002
[3] Salton G. The Smart Retrieval System-Experiments in Automatic Document Processing[M]. New Jersey,USA:Prenvtice Hall. Inl,1971
[4] 陈建超,郑启伦,李庆阳,等. 基于特征词关联性的同义词集挖掘算法[J]. 计算机应用研究,2009,26(7):2517-2519
[5] Schutze H,Pedersen J. A co-occurrence-based thesaurus and two applications to information retrieval[J]. Information Processing and Management,1997,33(3):307-318
[6] 吴云芳,石静,金彭. 基于图的同义词集自动获取方法[J]. 计算机研究与发展,2011,48(4):610-616
[7] Matsuo Y,Sakaki T,Uchiyama K,et al. Graph-based clustering using a Web search engine [C]// Proc of EMNLP. 2006:542-550
[8] Turney P D. Mining the Web for synonyms:PMI-IR versus LSA on TOEFL[C]// Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001). Freiburg,Germany, 2001:491-52
[9] 崔世起,刘群,林守勋,等. 中文缩略语自动抽取初探[C]// 孙茂松,陈群秀. 自然语言处理与大规模内容计算. 北京:清华大学出版社,2005:53-58
[10] 谢丽星,孙茂松,佟子健,等. 基于用户查询日志和锚文字的汉语缩略语识别[C]// 孙茂松,陈群秀. 中国计算语言学研究前沿进展. 北京:清华大学出版社,2009:551-556
[11] 田萱,杜小勇,李海华. 语义查询扩展中词语-概念相关度的计算[J]. 软件学报,2008,19(8):2043-2053
[12] 熊桂喜,王开锋. 基于语义的查询扩展研究[J]. 微计算机信息, 2008,24(30):177-178,187
[13] 杨清琳,李陶深,农健. 基于领域本体知识库的语义查询扩展[J]. 计算机工程与设计,2011,32(11):3853-3856
[14] 李海芳,史俊冰,段利国,等. 一种基于含糊同义词的查询扩展方法[J]. 计算机应用与软件,2011,28(12):41-43
[15] 余慧佳,刘奕群,张敏,等. 基于大规模日志分析的网络搜索引擎用户行为研究[J]. 中文信息学报,2007,21(1):109-114
[16] 岑荣伟,刘奕群,张敏,等. 基于日志挖掘的搜索引擎用户行为分析[J]. 中文信息学报,2010,24(3):49-54
[17] 窦志成,袁晓洁,何松柏. 大规模中文搜索日志中查询重复性分析[J]. 计算机工程,2008,34(21):40-41,44
[18] 张泽伟,矫健,张仰森. 基于 PMI-IR 的联想词表构造方法研究[J]. 计算机技术与发展,2014,24(6):140-144