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

曾安平[1,2] 李天瑞[1] 罗 川[1]

(西南交通大学信息科学与技术学院 成都 610031)[1] (宜宾学院计算机与信息工程学院 宜宾 644007)[2]

摘 要 在实际应用中,信息系统中数据的类型是多样的,它可能由类别型、数值型、模糊型等多种形式的数据组成。模糊粗糙集模型可以有效地解决多种类型数据共存情形下的信息处理问题。利用高斯核函数在数值和模糊数据非线性划分上的优势产生模糊关系可以较好地进行模糊粗糙数据分析。而实际的信息系统都是动态变化的,如何利用已有知识来增量更新模糊粗糙集模型的近似集问题是其应用于大数据处理的关键。针对该问题,讨论了模糊信息系统中对象集动态变化时近似集的更新原理,并提出了基于高斯核模糊粗糙集模型的近似集增量更新方法,最后通过实例验证了该方法的正确性和有效性。

关键词 模糊粗糙集,增量更新,高斯核函数

中图法分类号 TP391 **文献标识码** A

Incremental Approach for Updating Approximations of Gaussian Kernelized Fuzzy Rough Sets under Variation of Object Set

ZENG An-ping[1,2] LI Tian-rui[1] LUO Chuan[1]

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)[1]

(School of Computer and Information Engineering, Yibin University, Yibin 644007, China)[2]

Abstract In real-applications, there are many kinds of data in information systems. The data may consist of categorical, numerical, fuzzy values. Fuzzy rough set model can deal with this complex data. Gaussian kernels have been introduced to acquire fuzzy relations between samples described by fuzzy or numeric attributes to carry out fuzzy rough data analysis. In addition, the information systems often vary with time. How to use the previous knowledge to update approximations in fuzzy rough set model is a key step of its application on big data. This paper discussed the principles of updating approximations in fuzzy information systems under the variation of the object set. An approach for incrementally updating approximations of fuzzy rough set was then presented. Some examples were employed to illustrate the proposed approach.

Keywords Fuzzy rough set, Incremental updating, Gaussian kernel

1 引言

粗糙集理论是由 Pawlak 于 1982 年提出的一种建立在分类机制上的处理不确定性问题的数学工具,已被广泛用于人工智能、机器学习和数据挖掘等领域[1,2]。它将分类理解为在特定空间上的布尔等价关系,具有相同特征值的对象被认为是不可区分的,属于同一等价类。

经典的 Pawlak 粗糙集只能处理离散的类别型数据。而在实际的信息系统中可能存在各种类型的数据形式,如:类别型数据、数值型数据、模糊型数据等。为此,许多学者采用离散化的方法将相应的数据转换为类别型数据来进行处理,但离散化会造成大量的信息丢失,从而导致粗糙集模型处理结果的失真[3]。模糊粗糙集中对象间的模糊关系可以较好地处理类别、数值、模糊等多种类型共存的问题而不会引起信息的丢失。但模糊粗糙集有两个关键性的问题需要解决:(1)产生模糊等价关系并用该关系产生模糊粒;(2)如何利用模糊粒描述模糊概念。

模糊粗糙集是 Dubois 和 Prade 于 1990 年提出的,该模型要求模糊等价关系必须满足自反性、对称性和 max-min 传递性[4]。后来,Morsi 等引入 t-模和下半连续三角模 T 构成的模糊 T-相似关系对模糊粗糙集进行重新定义[5]。Yeung 等引入了一对 t-模 T 和 t-余模 S 对模糊粗糙集进行了重新定义,提出了更一般化的模糊粗糙集模型[6]。胡清华等就如何从数据中有效地产生模糊相似关系进行了系统的讨论并提出了基于高斯核的模糊粗糙集,其首次将满足自反性、对称性和 T_{cos}-传递性的高斯核函数引入到模糊等价关系的产生上来[7]。他们还提出核化的模糊粗糙集,将高斯核函数推广到其它核函数,从而给出了系统的等价类产生方法[8]。近年来,

到稿日期:2012-09-16 返修日期:2013-01-08 本文受国家自然科学基金项目(61175047,U1230117),四川省教育厅青年基金项目(13ZB0210),西南交通大学博士生创新基金项目(2013)资助。

曾安平(1975—),男,博士生,讲师,CCF 会员,主要研究方向为数据挖掘与知识发现,E-mail:zengap@126.com;李天瑞(1969—),男,博士,教授,博士生导师,主要研究方向为数据挖掘与知识发现、粒计算与粗糙集、云计算。

模糊粗糙集在特征评价、属性约简、规则提取、医学分析和股票预测等方面得到了成功应用[9-11]。

传统的基于粗糙集和模糊粗糙集等模型的近似集更新方法是针对静态信息系统的，在面对大规模动态复杂现实问题时都面临着高复杂性计算的困难。增量式近似集更新方法由于能够充分利用已有的近似集知识，因此能有效提高近似集求解问题的效率[12]。故，很多学者致力于利用增量式近似集更新方法来提高基于粗糙集理论的知识发现方法的效率。本文主要研究高斯核模糊粗糙集模型中增量式近似集更新方法。信息系统动态变化主要可以从属性值粗化细化、属性集增删、对象集增删以及它们同时变化 4 个方面来进行考虑。本文利用模糊粗糙集在处理不确定性问题的优势，讨论在对象集不断变化的模糊信息系统(Fuzzy Information System, FIS)中近似集的增量更新理论，这对于将粗糙集拓展用于大数据处理领域有着十分重要的理论价值和现实意义。

2 模糊粗糙集

2.1 粗糙集

粗糙集理论将分类理解为在特定空间上的等价关系，而等价关系则构成了对该空间的划分。它可以通过已定义好的一对精确概念：下近似集和上近似集将不精确或不确定的知识进行近似的表示。

定义 1[1] 给定 Pawlak 近似空间(U,R)，论域U非空，$R\subseteq U\times U$是论域U上的一个等价关系。U/R代表基于等价关系R的划分，对象$x\in U$，$[x]_R$表示x所在的等价类。对任意概念$X\subseteq U$，其下近似集和上近似集定义如下：

$$\underline{RX}=\{x\in U|[x]_R\subseteq X\}$$
$$\overline{RX}=\{x\in U|[x]_R\cap X\neq\phi\} \tag{1}$$

2.2 模糊集

定义 2[13] 设U为论域，则称由如下实值函数

$$\mu_A:\begin{matrix}U\to[0,1]\\ u\mapsto\mu_A(u)\end{matrix} \tag{2}$$

所确定的集合A为U上的模糊集合。$\mu_A(u)$称为u对于A的隶属度，有时简称$A(u)$。

模糊集具有两种表示方法：

(1)Zadeh 表示法

当论域$U=\{x_1,x_2,\cdots,x_n\}$时，模糊集合A可表示为：$A=A(x_1)/x_1+A(x_2)/x_2+\cdots+A(x_n)/x_n$。

例 1 有 5 个科研项目$U=\{x_1,x_2,x_3,x_4,x_5\}$，专家用百分制打分，分别为 87,73,94,85,79，用“成果优秀”这一模糊概念来代替百分制成绩，则 5 个项目构成的模糊成绩集合$A=0.87/x_1+0.73/x_2+0.94/x_3+0.85/x_4+0.79/x_5$。

(2)向量表示法

$A=\langle A(x_1),A(x_2),\cdots,A(x_n)\rangle$或$A=\langle A(x_1)/u_1,A(x_2)/x_2,\cdots,A(x_n)/x_n\rangle$。

例 1 中模糊集A表示为

$A=\langle 0.87,0.73,0.94,0.85,0.79\rangle$或$A=\langle 0.87/x_1,0.73/x_2,0.94/x_3,0.85/x_4,0.79/x_5\rangle$

本文将采用向量表示法表示模糊集，为了和粗糙集中上下近似集的表示方法统一，下文采用大括号“{ }”代替“〈〉”。

定义 3[13] 设U,V为两个论域，若$R\in F(U\times V)$，则称R为U到V的一个模糊关系。若$(u\times v)\in(U\times V)$，则称$u$对$v$具有关系$R$的相关程度。

定义 4[13] 设$R\in F(U\times U)$，若R满足

(1)自反性：$R(x,x)=1$；

(2)对称性：$R(x,y)=R(y,x)$；

(3)传递性：$\min(R(x,y),R(y,z))\leqslant R(x,z)$。

则称R为模糊等价关系。

定义 5[14] 记$I=[0,1]$，二元函数$T:I\times I\to I$，$a,b,c,d\in I$，若满足下列条件：

(1)两极律：$T(a,1)=a$；

(2)交换律：$T(a,b)=T(b,a)$；

(3)结合律：$T(T(a,b),c)=T(a,T(b,c))$；

(4)单调律：$a\leqslant c,b\leqslant d\Rightarrow T(a,b)\leqslant T(c,d)$。

则称T为I上的三角模或T模。

min 是一个特殊的三角模，式(3)也是一个三角模[7]。

$$T_{\cos}(a,b)=\max\{ab-\sqrt{(1-a^2)(1-b^2)},0\} \tag{3}$$

本文讨论的模糊集以及模糊粗糙集是基于$T_{\cos}$模的。

定义 6[5] 设T是某三角模，如果R满足自反性：$R(x,y)=R(y,x)$，对称性：$R(x,x)=1$和T传递性：$T(R(x,y),R(y,z))\leqslant R(x,z)$，则$R$是一个$T$模糊等价关系。

显然，如果T是$T_{\cos}$模，则根据定义 6，如果R同时满足自反性、对称性和$T_{\cos}$传递性，则R是一个$T_{\cos}$模糊等价关系。

2.3 基于高斯核函数的模糊集

目前，高斯核函数被作为分类核广泛用于 SVM 和 RBF 人工神经网络等领域[15]。在高维特征空间中，利用高斯核函数进行非线性问题求解具有较好的性能和计算效率。

设U为非空有限的论域，对象$x_i\in U$，以n维向量形式表示为$x_i=\langle x_{i1},x_{i2},\cdots,x_{in}\rangle\in R^n$，$U$可以看作是$R^n$的子集(在不引起歧义的情况下，下文将用$R$代表$R^n$)。任意两个对象之间的相似度可以通过高斯核函数

$$k(x_i,x_j)=\exp\left(\frac{-\|x_i-x_j\|^2}{2\delta^2}\right) \tag{4}$$

来进行计算，其中$\|x_i-x_j\|$是x_i和x_j之间的欧氏距离。由此可得：

(1)$k(x_i,x_j)\in[0,1]$；

(2)$k(x_i,x_j)=k(x_j,x_i)$；

(3)$k(x_i,x_i)=1$。

定理 1[7] 任何核函数$k:U\times U\to[0,1]$，如果$k(x,x)=1$，则该核函数满足$T_{\cos}$传递性且

$$T_{\cos}(a,b)=\max\{ab-\sqrt{(1-a^2)(1-b^2)},0\}$$

如果一个关系同时满足自反性、对称性和传递性，则该关系是等价关系，很容易得出以下推论。

推论 1[7] 通过高斯核函数计算的模糊关系R_G是一个$T_{\cos}$模糊等价关系。

例 2 在例 1 基础上，如果 5 个科研项目有两个专家用百分制打分，专家a的打分为 87,73,94,85,79；专家b的打分为 89,80,85,87,65；综合成绩(决策属性)分别为“优秀”，“一般”，“优秀”，“优秀”，“一般”。打分分别用“成果优秀”这一模糊概念来表示，见表 1。

表1　5个科研项目成绩表

项目(x)	专家(a)	专家(b)	综合成绩(d)
x_1	0.87	0.89	优秀
x_2	0.73	0.80	一般
x_3	0.94	0.85	优秀
x_4	0.85	0.87	优秀
x_5	0.79	0.65	一般

显然，只考虑条件属性(专家 a 和专家 b)，令 $\delta^2=1$，我们可以求出各对象间的高斯核相似度，即可计算出所有对象之间的相似度，从而得到如下关系矩阵：

$$R_G=\begin{pmatrix}1 & 0.9203 & 0.9605 & 0.9860 & 0.8812\\ 0.9203 & 1 & 0.8977 & 0.9180 & 0.9223\\ 0.9605 & 0.8977 & 1 & 0.9461 & 0.8824\\ 0.9860 & 0.9180 & 0.9461 & 1 & 0.8922\\ 0.8812 & 0.9223 & 0.8824 & 0.8922 & 1\end{pmatrix}$$

从矩阵中可以看出，通过高斯核计算的相似关系 R_G 满足对称性和自反性，而且可以验证 R_G 满足 $T_{\cos}$ 传递性。因此，R_G 是一个 $T_{\cos}$-模糊等价关系。

2.4　基于高斯核函数的模糊粗糙集

首先来看一般意义下的模糊粗糙集的定义。

定义 7[4]　设 R 是论域 U 上的模糊等价关系，X 是 U 上的模糊子集，$x\in X$，其隶属函数记为 $X(x)$，则 X 的下近似集和上近似集也是一对 U 的模糊子集，且隶属函数分别为：

$$\begin{cases}\underline{R_{\max}}X(x)=\inf\limits_{y\in U}\max(1-R(x,y),X(y))\\ \overline{R_{\min}}X(x)=\sup\limits_{y\in U}\min(R(x,y),X(y))\end{cases}\tag{5}$$

min 是特殊的 t-模 T，而 max 是特殊的 t-余模 S。为了统一模糊粗糙集的定义，文献[6]引入了一对 t-模 T 和 t-余模 S 对模糊粗糙集进行重新定义。

定义 8[6]　给定论域 U 中的 T-模糊等价关系 R，模糊概念 X 的模糊下近似和模糊上近似定义如下：

$$\begin{cases}\underline{R_S}X(x)=\inf\limits_{y\in U}S(N(R(x,y)),X(y))\\ \overline{R_T}X(x)=\sup\limits_{y\in U}T(R(x,y),X(y))\end{cases}\tag{6}$$

如果将定义 8 中的 T-模糊等价关系 R 考虑为基于高斯核的 $T_{\cos}$-模糊等价关系 R_G，则可以得出基于高斯核的模糊粗糙集定义。

定义 9[7]　给定论域 U 中基于高斯核函数的 $T_{\cos}$-模糊等价关系 R_G，模糊概念 X 的模糊下近似和模糊上近似定义如下：

$$\begin{cases}\underline{R_{G\theta}}X(x)=\inf\limits_{y\in U}\theta_{\cos}(R_G(x,y),X(y))\\ \overline{R_{GT}}X(x)=\sup\limits_{y\in U}T_{\cos}(R_G(x,y),X(y))\end{cases}\tag{7}$$

其中

$$\theta_{\cos}(a,b)=\begin{cases}1, & a\leqslant b\\ ab+\sqrt{(1-a^2)(1-b^2)}, & a>b\end{cases}\tag{8}$$

假设决策属性 d 可以将论域 U 划分为 $\{d_1,d_2,\cdots,d_l\}$。在实际运算过程中，可作如下处理：$\forall x\in d_i$，则 $d_i(x)=1$，否则 $d_i(x)=0$。由此，可将决策类转换为模糊隶属度的方式来处理，从而得到如下性质：

性质 1[7]　对于任意决策类 d_i，其下近似与上近似如下：

$$\begin{cases}\underline{R_{G\theta}}d_i(x)=\inf\limits_{y\notin d_i}(\sqrt{1-R_G^2(x,y)})\\ \overline{R_{GT}}d_i(x)=\sup\limits_{y\in d_i}R_G(x,y)\end{cases}\tag{9}$$

由性质 1 可以看出，任意对象 x 在决策类 d_i 的下近似集中的隶属度是由不属于类 d_i 中的与 x 最近的对象所决定的，而上近似集中的隶属度是由属于类 d_i 中的与 x 最近的对象所决定的。性质在文献[7]中进行了证明，这里证明略。

例 3　在例 2 的基础上，求综合成绩为“优秀”的对象的下近似集与上近似集。

按上述思想，综合成绩(d)被划分为“优秀”(d_1)和“一般”(d_2)两个决策类，其中 $d_1=\{x_1,x_3,x_4\}$。

根据性质 1 可得决策类 d_1 的下近似集和上近似集：

$$\begin{aligned}\underline{R_{G\theta}}d_1(x_1)&=\inf_{y\notin d_1}(\sqrt{1-R_G^2(x_1,y)})\\ &=\inf_{y\in\{x_2,x_5\}}(\sqrt{1-R_G^2(x_1,y)})\\ &=\min\{\sqrt{1-0.9203^2},\sqrt{1-0.8812^2}\}\\ &=0.39\end{aligned}$$

$\underline{R_{G\theta}}d_1(x_2)=0$，$\underline{R_{G\theta}}d_1(x_3)=0.44$

$\underline{R_{G\theta}}d_1(x_4)=0.397$，$\underline{R_{G\theta}}d_1(x_5)=0$

$$\begin{aligned}\overline{R_{GT}}d_1(x_1)&=\sup_{y\in d_1}R_G(x,y)=\sup_{y\in\{x_1,x_3,x_4\}}R_G(x,y)\\ &=\max\{1,0.9605,0.986\}=1\end{aligned}$$

同理可得：

$\overline{R_{GT}}d_1(x_2)=0.918$，$\overline{R_{GT}}d_1(x_3)=1$

$\overline{R_{GT}}d_1(x_4)=1$，$\overline{R_{GT}}d_1(x_5)=0.8922$

因此，下近似集和上近似集分别为：

$\underline{R_{G\theta}}d_1=\{0.39/x_1,0.44/x_3,0.397/x_4\}$

$\overline{R_{GT}}d_1=\{1/x_1,0.918/x_2,1/x_3,1/x_4,0.8922/x_5\}$

注意：为了简化表达式，在上下近似集中如果某个对象的隶属度为 0，则将该对象从上下近似集中去除。

3　对象变化下的增量更新近似集方法

给定一个 t 时刻的模糊信息系统 $FIS=(U;C\cup D;V;f)$，$U\neq\emptyset$，对任意模糊子集 $X\subseteq U$，其下近似和上近似集分别表示为 $\underline{R_{G\theta}^{(t)}}X$，$\overline{R_{GT}^{(t)}}X$。在对象变化时，设 x^+ 表示在 $t+1$ 时刻进入系统中的对象，x^- 表示在 $t+1$ 时刻从系统中删除的对象。对象发生变化后的模糊信息系统记为 $FIS'=(U';C'\cup D';V';f')$。$t+1$ 时刻 X 的下近似和上近似集分别表示为 $\underline{R_{G\theta}^{(t+1)}}X$，$\overline{R_{GT}^{(t+1)}}X$。

3.1　对象增加的情形

假定在 $t+1$ 时刻对象 x^+ 进入 FIS，则 $U'=U\cup\{x^+\}$。首先根据高斯核函数计算 x^+ 与论域中所有对象的相似度，有以下性质成立。

性质 2　当某对象 x^+ 进入论域 U，则有：

(1)如果 $\exists x\in U$，使得 $f(x,d)=f(x^+,d)$，则 x^+ 的进入不会产生新的决策类；

(2)如果 $\forall x\in U$，均有 $f(x,d)\neq f(x^+,d)$，则 x^+ 的进入将产生新的决策类。

根据性质 2 可判定 x^+ 是否产生新的决策类。

性质 3　$\forall d_i\in U'/D'$，$\forall x\in U'$，如果 x^+ 使得产生新决策类，令新类为 d_{l+1}，则

$$\underline{R_{G\theta}^{(t+1)}}d_i(x)=\begin{cases}\inf\{\underline{R_{G\theta}^{(t)}}d_i(x)\sqrt{1-R_G^2(x,x^+)}\}, & x\neq x^+\\ \inf\limits_{y\notin d_i}(\sqrt{1-R_G^2(x,y)}\}, & \text{其它}\end{cases}\tag{10}$$

$$\overline{R_{GT}^{(t+1)}}d_i(x)=\begin{cases}\overline{R_{GT}^{(t)}}d_i(x), & x\neq x^+\\ \sup\limits_{y\in d_i}R_G(x,y), & \text{其它}\end{cases}\tag{11}$$

证明：对于决策类 d_i，由于 x^+ 要产生新决策类，因此 $x^+\notin d_i$，根据性质 1，对于 $\forall x\neq x^+$，

$$\begin{aligned}\underline{R_{GB}^{(t+1)}}d_i(x)&=\inf_{y\notin d_i}(\sqrt{1-R_G^2(x,y)})\\&=\inf_{y\notin d_i\text{ and }y\neq x^+}(\sqrt{1-R_G^2(x,y)})\wedge\inf_{y=x^+}(\sqrt{1-R_G^2(x,y)})\\&=\underline{R_{GB}^{(t)}}d_i(x)\wedge\inf_{y=x^+}(\sqrt{1-R_G^2(x,y)})\\&=\inf(\underline{R_{GB}^{(t)}}d_i(x),\sqrt{1-R_G^2(x,y)})\end{aligned}$$

对于上近似集，根据性质 1，很容易得证。

对于 $x=x^+$，显然需要根据性质 1 计算上下近似。

性质 4 $\forall d_i\in U'/D'$，$\forall x\in U'$，如果 x^+ 使得不产生新决策类，且 $x^+\in d_k$，则

$$\underline{R_{GB}^{(t+1)}}d_i(x)=\begin{cases}\underline{R_{GB}^{(t)}}d_i(x), & d_i\neq d_k\text{ and }x\neq x^+\\ \inf\{\underline{R_{GB}^{(t)}}d_i(x),\sqrt{1-R_G(x,x^+)}\}, & d_i\neq d_k\text{ and }x\neq x^+\\ \inf\limits_{y\notin d_i}(\sqrt{1-R_G(x,y)}), & \text{其它}\end{cases}\tag{12}$$

$$\overline{R_{GT}^{(t+1)}}d_i(x)=\begin{cases}\overline{R_{GT}^{(t)}}d_i(x), & d_i\neq d_k\text{ and }x\neq x^+\\ \sup\{\overline{R_{GT}^{(t)}}d_i(x),R_G(x,x^+)\}, & d_i=d_k\text{ and }x\neq x^+\\ \sup\limits_{y\in d_k}R_G(x,y), & \text{其它}\end{cases}\tag{13}$$

可知，通过性质 3、性质 4 可计算相应的上下近似集。

证明：当 $d_i=d_k$ and $x\neq x^+$ 时，根据性质 1，由于 d_i 的下近似集由不属于 d_i 的对象决定，因此，$t+1$ 时刻和 t 时刻的下近似集相同。而上近似集

$$\begin{aligned}\overline{R_{GT}^{(t+1)}}d_i(x)&=\sup_{y\in d_i}R_G(x,y)\\&=\sup_{y\in d_i\text{ and }x\neq x^+}R_G(x,y)\vee R_G(x,x^+)\\&=\sup\{\overline{R_{GT}^{(t)}}d_i(x),R_G(x,x^+)\}\end{aligned}$$

当 $d_i\neq d_k$ and $x\neq x^+$ 时，

$$\begin{aligned}\underline{R_{GB}^{(t+1)}}d_i(x)&=\inf_{y\notin d_i}(\sqrt{1-R_G^2(x,y)})\\&=\inf_{y\notin d_i\text{ and }x\neq x^+}(\sqrt{1-R_G^2(x,y)})\wedge\sqrt{1-R_G^2(x,x^+)}\\&=\inf\{\underline{R_{GB}^{(t)}}d_i(x),\sqrt{1-R_G^2(x,x^+)}\}\end{aligned}$$

由于 d_i 的上近似集由属于 d_i 的对象决定，因此，$t+1$ 时刻和 t 时刻的上近似集相同。

对于 $x=x^+$，显然需要根据性质 1 计算上下近似。

例 4　在表 1 基础上，假定项目 $x_6=(0.58,0.64,$差$)$和 $x_7=(0.7,0.7,$优秀$)$进入系统中(见表 2)。

(1)首先按照式(4)计算 x_6 和其它对象的高斯核相似关系：

$k(x_1,x_6)=0.82$，$k(x_2,x_6)=0.896$，$k(x_3,x_6)=0.8055$，$k(x_4,x_6)=0.829$，$k(x_5,x_6)=0.8986$。

表 2　对象增加情形表

项目(x)	专家(a)	专家(b)	综合成绩(d)
x_1	0.87	0.89	优秀
x_2	0.73	0.80	一般
x_3	0.94	0.85	优秀
x_4	0.85	0.87	优秀
x_5	0.79	0.65	一般
$-->x_6$	0.58	0.64	差
$-->x_7$	0.70	0.70	优秀

根据性质 2，由于 x_6 的综合成绩为“差”，因此产生了新的决策类 d_3。由性质 3 的式(10)知，下近似集

$$\underline{R_{GB}^{(t+1)}}d_1(x_1)=\inf\{\underline{R_{GB}^{(t)}}d_1(x_1),\sqrt{1-R_G^2(x_1,x_6)}\}=\inf\{0.39,0.57\}=0.39$$

同理

$\underline{R_{GB}^{(t+1)}}d_1(x_2)=0$，$\underline{R_{GB}^{(t+1)}}d_1(x_3)=0.44$

$\underline{R_{GB}^{(t+1)}}d_1(x_4)=0.397$，$\underline{R_{GB}^{(t+1)}}d_1(x_5)=0$

$$\begin{aligned}\underline{R_{GB}^{(t+1)}}d_1(x_6)&=\inf_{y\notin d_1}(\sqrt{1-R_G^2(x_6,y)})\\&=\inf_{y\in\{x_2,x_5,x_6\}}(\sqrt{1-R_G^2(x_6,y)})\\&=\min\{\sqrt{1-0.896^2},\sqrt{1-0.8986^2},\sqrt{1-1^2}\}\\&=0\end{aligned}$$

由性质 3 的式(11)，有上近似集

$\overline{R_{GT}^{(t+1)}}d_1(x_1)=\overline{R_{GT}^{(t)}}d_1(x_1)=1$

同理

$\overline{R_{GT}^{(t+1)}}d_1(x_2)=0.918$，$\overline{R_{GT}^{(t+1)}}d_1(x_3)=1$

$\overline{R_{GT}^{(t+1)}}d_1(x_4)=1$，$\overline{R_{GT}^{(t+1)}}d_1(x_5)=0.8922$

$$\overline{R_{GT}^{(t+1)}}d_1(x_6)=\sup_{y\in d_1}R_G(x_6,y)=\sup_{y\in\{x_1,x_3,x_4\}}R_G(x_6,y)=0.829$$

因此，x_6 对象增加后下近似集和上近似集为：

$\underline{R_{GB}^{(t+1)}}d_1=\{0.39/x_1,0.44/x_3,0.397/x_4\}$

$\overline{R_{GT}}d_1=\{1/x_1,0.918/x_2,1/x_3,1/x_4,0.8922/x_5,0.829/x_6\}$

如果需要计算 d_3 的下近似集和上近似集，则根据性质 3 和性质 4 也可以很方便求得，这里略。

(2)按照式(4)计算 x_7 和其它对象的高斯相似关系：

$k(x_1,x_7)=0.88$，$k(x_2,x_7)=0.949$

$k(x_3,x_7)=0.868$，$k(x_4,x_7)=0.893$

$k(x_5,x_7)=0.95$，$k(x_6,x_7)=0.935$

根据性质 2，由于 x_7 的综合成绩为“优秀”，因此 $x_7\in d_1$。而 $x_1,x_3,x_4\in d_1$，根据性质 4 的式(12)，下近似集有

$\underline{R_{GB}^{(t+1)}}d_1(x_1)=\underline{R_{GB}^{(t)}}d_1(x_1)=0.39$

同理

$\underline{R_{GB}^{(t+1)}}d_1(x_3)=0.44$，$\underline{R_{GB}^{(t+1)}}d_1(x_4)=0.397$

对于 $x_2,x_5\notin d_1$ 且不等于 x_7，

$$\underline{R_{GB}^{(t+1)}}d_1(x_2)=\inf\{\underline{R_{GB}^{(t)}}d_1(x_1),\sqrt{1-R_G^2(x_2,x_7)}\}=\inf\{0,\sqrt{1-R_G^2(x_2,x_7)}\}=0$$

同理

$\underline{R_{GB}^{(t+1)}}d_1(x_5)=0$，$\underline{R_{GB}^{(t+1)}}d_1(x_6)=0$

对于 x_7，

$$\begin{aligned}\underline{R_{GB}^{(t+1)}}d_1(x_7)&=\inf_{y\notin d_1}\{\sqrt{1-R_G^2(x_7,y)}\}\\&=\inf_{y\in\{x_2,x_5,x_6\}}\{\sqrt{1-R_G^2(x_7,y)}\}\\&=0.312\end{aligned}$$

对于上近似集，因为 x_1, x_3, x_4 和 x_7 属于同一决策类，根据性质 4 的式(13)，有

$\overline{R_{GT}^{(t+1)}}d_i(x_1)=\sup\{\overline{R_{GT}^{(t)}}d_i(x_1), R_G(x,x^+)\}=1$

同理

$\overline{R_{GT}^{(t+1)}}d_1(x_3)=1, \overline{R_{GT}^{(t+1)}}d_1(x_4)=1$

对于 x_2, x_5, x_6 和 x_7 不属于同一决策类，根据性质 4，上、下近似中的隶属度保持不变：

$\underline{R_{GT}}d_1(x_2)=0.918, \underline{R_{GT}}d_1(x_5)=0.8922$

$\overline{R_{GT}}d_1(x_2)=0.829$

另外，根据性质 4 有

$\overline{R_{GT}^{(t+1)}}d_i(x_7)=\sup_{y\in d_k} R_G(x,y)=\sup_{y\in\{x_1,x_3,x_4,x_7\}} R_G(x,y)=1$

因此，x_7 对象增加后下近似集和上近似集为：

$\underline{R_{GT}^{(t+1)}}d_1=\{0.39/x_1, 0.44/x_3, 0.397/x_4, 0.312/x_7\}$

$\overline{R_{GT}}d_1=\{1/x_1, 0.918/x_2, 1/x_3, 1/x_4, 0.8922/x_5, 0.829/x_6, 1/x_7\}$

从上面两个性质可看出，当对象增加时，$t+1$ 时刻的上下近似集只需在上一时刻(t 时刻)的近似集基础上进行少量的变动计算即可。因此，根据性质 3、性质 4 可以完成增加对象情况下的增量式更新近似集。

3.2 对象删除的情形

假定在 $t+1$ 时刻对象 x^- 移出 FIS，则 $U'=U-\{x\}$。首先根据性质 5 判断 x^- 的删除是否会引起决策类的移除。然后根据性质 6、性质 7 调整各对象的上、下近似集。

性质 5 在 $t+1$ 时刻，某对象 x^- 移出论域，d 为决策属性，则有：

(1)如果 $\exists x\in U'$，使得 $f(x,d)=f(x^-,d)$，则 x^- 的删除不会引起决策类的移除；

(2)如果 $\forall x\in U'$，均有 $f(x,d)\neq f(x^-,d)$，则 x^- 的删除将引起决策类的删除。

当决策类变化确定后，根据性质 6、性质 7 可以计算各决策类的上、下近似集。

性质 6 $\forall d_i\in U'/D', \forall x\in U'$，如果 x^- 使得某决策类被删除，令该类为 d_k，则

$$\underline{R_{GT}^{(t+1)}}d_i(x)=\begin{cases}\underline{R_{GT}^{(t+1)}}d_i(x), & \underline{R_{GT}^{(t+1)}}d_i(x)<\sqrt{1-R_G^2(x,x^-)}\\ \inf\limits_{y\in d_i}\sqrt{1-R_G^2(x,x^-)}, & \text{其它}\end{cases} \tag{14}$$

$$\overline{R_{GT}^{(t+1)}}d_i(x)=\overline{R_{GT}^{(t)}}d_i(x) \tag{15}$$

证明：当某类决策类被删除时，对于其它决策类 d_i，若删除对象即是与对象 x 最近的对象，则 x 的下近似需要重新计算，否则保持不变。

对于 d_i 的上近似集，由于只与属于决策类 d_i 的对象有关，因此，删除不属于 d_i 的对象对上近似集不会发生影响。

性质 7 $\forall d_i\in U'/D', \forall x\in U'$，如果 x^- 不能使某决策类被删除，且 $x^-\in d_k$，则

$$\underline{R_{GT}^{(t+1)}}d_i(x)=\begin{cases}\underline{R_{GT}^{(t)}}d_i(x), & d_i=d_k\\ \underline{R_{GT}^{(t)}}d_i(x), & d_i=d_k \text{ and } \underline{R_{GT}^{(t)}}d_i(x)<\sqrt{1-R_G(x,x^-)}\\ \inf\limits_{y\in d_k}\sqrt{1-R_G(x,x^-)}, & \text{其它}\end{cases} \tag{16}$$

$$\overline{R_{GT}^{(t+1)}}d_i(x)=\begin{cases}\overline{R_{GT}^{(t)}}d_i(x), & d_i\neq d_k\\ \overline{R_{GT}^{(t)}}d_i(x), & d_i=d_k \text{ and } \overline{R_{GT}^{(t)}}d_i(x)>R_G(x,y)\\ \sup\limits_{y\in d_k}R_G(x,y), & \text{其它}\end{cases} \tag{17}$$

证明：当 $d_i=d_k$ 时，由于 d_i 下近似集由不属于 d_i 的对象决定，且 $x^-\in d_i$，显然下近似不会发生变化。而上近似中，如果 x^- 就是与对象 x 最近的对象，则上近似集需要重新计算，否则上近似集不会发生变化。

当 $d_i\neq d_k$ 时，如果删除对象 x^- 就是与对象 x 最近的对象，则 x 的下近似需要重新计算，否则不用计算。而对于上近似，由于 d_i 上近似由属于 d_i 的对象决定，且 $x^-\notin d_i$，因此显然上近似不会发生变化。

例 5 在表 1 基础上删除对象 x_1(见表 3)。

表 3 对象删除情形

项目(x)	专家(a)	专家(b)	综合成绩(d)
x_1	~~0.87~~	~~0.89~~	~~优秀~~
x_2	0.73	0.80	一般
x_3	0.94	0.85	优秀
x_4	0.85	0.87	优秀
x_5	0.79	0.65	一般

从例 3 中可知：

$\underline{R_{GT}}d_1=\{0.39/x_1, 0.44/x_3, 0.397/x_4\}$

$\overline{R_{GT}}d_1=\{1/x_1, 0.918/x_2, 1/x_3, 1/x_4, 0.8922/x_5\}$

从表 3 中可以看出，删除 x_1 不会删除决策类，且 $x_1\in d_1$。

(1)对于下近似集，$x_3, x_4\in d_1$，根据性质 7 的式(16)有：

$\underline{R_{GT}^{(t+1)}}d_1(x_3)=\underline{R_{GT}^{(t)}}d_1(x_3)=0.44$

$\underline{R_{GT}^{(t+1)}}d_1(x_4)=\underline{R_{GT}^{(t)}}d_1(x_4)=0.397$

由于 $x_2, x_5\notin d_1$，前面已经得出 $\underline{R_{GT}^{(t)}}d_1(x_2)=0$，$\underline{R_{GT}^{(t)}}d_1(x_5)=0$，已经是最小，根据式(16)：$\underline{R_{GT}^{(t+1)}}d_1(x_2)=0$，$\underline{R_{GT}^{(t+1)}}d_1(x_5)=0$。

(2)对于上近似集，$x_3, x_4\in d_1$，$\overline{R_{GT}^{(t)}}d_1(x_3)=1$，$\overline{R_{GT}^{(t)}}d_1(x_4)=1$ 已经是最大隶属度，根据性质 7 的式(17)有：

$\overline{R_{GT}^{(t+1)}}d_1(x_3)=1, \overline{R_{GT}^{(t+1)}}d_1(x_4)=1$

对于 $x_2, x_5\notin d_1$，根据式(17)，显然隶属度保持不变。

因此，x_1 对象删除后下近似集和上近似集为：

$\underline{R_{GT}^{(t+1)}}d_1=\{0.44/x_3, 0.397/x_4\}$

$\overline{R_{GT}}d_1=\{0.918/x_2, 1/x_3, 1/x_4, 0.8922/x_5\}$

从性质 6、性质 7 可以得出结论，当对象删除时，$t+1$ 时刻的上下近似集计算只需要在上一时刻 t 的上下近似集基础上计算少量变动情况即可。因此，根据性质 6、性质 7 可以完成删除对象情况下的增量式更新近似集。

结束语 在实际的信息系统中，对象通常是动态变化的，即动态增加或者删除，本文讨论了这种动态变化情形下模糊粗糙集模型中近似集的增量更新原理，给出了相应的上下近似集更新方法，并通过相应的实例分析验证了所提方法。如何根据特定的应用场景确定高斯核阈值、利用 UCI 数据集测试增量方法的性能，以及将该方法应用于特征提取等，将是我们今后的研究工作重点。

(下转第 215 页)

方面将扩大树库的规模，比如把宾州汉语树库通过结构转换增加到清华汉语树库中，同时，把本文获取类别知识的方法应用于以印欧语言为代表的英语当中；另一个方面把通过聚类方法挖掘到的类别知识应用到具体的句法结构消歧中，具体通过该方法获取的词汇类别知识的优越性加以验证。

参 考 文 献

[1] Aarts F. On the distribution of noun-phrase types in English clause-structue [J]. Lingua, 1971(26): 281-293

[2] Haan P D. Postmodifying clauses in the English Noun Phrase: a corpus-based study [M]. Amsterdam: Rodopi, 1989

[3] Maestre M D L. Noun Phrasecom Plexityasasty lemarker: anexe-reiseinstylisti analysis [J]. Atlantis, 1998(2): 91-105

[4] Maienborn C. On the Position and Interpretation of Locative Modifiers [J]. Natural Language Semantics, 2001, 9(2): 191-240

[5] Uzuner, Katz B. A Comparative Study of Language Models for Book and Author Recognition [J]. Lecture Notes in Computer Science, 2005(3651): 969-980

[6] Kalampakas A. The Syntactic Complexity of Eulerian Graphs [J]. Lecture Notes in Computer Science, 2007(4728): 208-217

[7] Stepanov A, Tsa W D. Cartography and licensing of wh-adjuncts: a cross-linguistic perspective [J]. Natural Language & Linguistic Theory, 2008, 26(3): 589-638

[8] 陈小荷. 从自动句法分析角度看汉语词类问题[J]. 语言教学与研究, 1999(03): 72

[9] 徐艳华. 现代汉语实词语法功能考察及词类体系重构[D]. 南京: 南京师范大学, 2006

[10] Boley D, et al. Partitioning-Based clustering for web document categorization[J]. Decision Support System Journal, 1999, 27(3): 329-341

[11] Mao J, Jain A K. A self-organizing network for hyperellipsoidal clustering [J]. IEEE Trans. Neural Networks, 1996, 7(2): 16-29

[12] Cai W L, Chen S C, Zhang D Q. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation[J]. Pattern Recognition, 2007, 40(3): 825-833

[13] 崔尚卿. 基于不均匀密度的自动聚类算法[J]. 计算机工程, 2008(23): 86-88

[14] 王伟. 文本自动聚类技术研究[J]. 情报杂志, 2009(02): 94-96

[15] 王舵. 一种快速词自动聚类算法[J]. 计算机应用与软件, 2010(08): 277-278

[16] 潘章明. 半监督的自动聚类[J]. 计算机应用, 2010(03): 2614-2616

[17] 于洪. 一种基于决策粗糙集的自动聚类方法[J]. 计算机科学, 2011(1): 221-224

(上接第 177 页)

参 考 文 献

[1] Zdzislaw P. Rough Sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356

[2] Zdzislaw P. Why Rough Sets? [A]// The Fifth IEEE International Conference on Fuzzy Systems[C]. Louisiana, New Orleans: IEEE Press, 1996: 738-743

[3] Richard J, Shen Qiang. Fuzzy-Rough Sets Assisted Attribute Selection[J]. IEEE Transactions on Fuzzy Systems, 2007, 15(1): 73-89

[4] Didier D, Henri P. Rough Fuzzy Sets and Fuzzy Rough Sets[J]. International Journal of General Systems, 1990, 17(2/3): 191-209

[5] Nehad M, Yakout M. Axiomatics for Fuzzy Rough Set[J]. Fuzzy Sets System, 1998, 100(1-3): 327-342

[6] So Y D, Chen De-gang, Eric T C C, et al. On the Generalization of Fuzzy Rough Sets[J]. IEEE Transactions on Fuzzy System, 2005, 13: 343-361

[7] Hu Qing-hua, Zhang Lei, Chen De-gang, et al. Gaussian Kernel based Fuzzy rough Sets: Model, Uncertainty Measures and Applications[J]. International Journal of Approximate Reasoning, 2010, 51: 453-471

[8] Hu Qing-hua, Yu Da-ren, Pedrycz W, et al. Kernelized Fuzzy Rough Sets and Their Applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(11): 1649-1667

[9] Hong T-P, Wang T-T, Wang S-L, et al. Learning a Coverage Set of Maximally General Fuzzy Rules by Rough Sets[J]. Expert Systems with Applications, 2000, 19(2): 97-103

[10] Tsai Y-C, Cheng C-H, Chang Jing-rong. Entropy-Based Fuzzy Rough Classification Approach for Extracting Classification Rules[J]. Expert Systems with Applications, 2006, 31(2): 436-443

[11] Wang Xi-zhao, Tsang E C C, Zhao Su-yun, et al. Learning Fuzzy Rules from Fuzzy Samples Based on Rough Set Technique[J]. Information Sciences, 2007, 177(20): 4493-4514

[12] Li Tian-rui, Ruan Da, Greet W, et al. A rough Sets based Characteristic Relation Approach for Dynamic Attribute Generalization in Data Mining[J]. Knowledge-Based Systems, 2007, 20(5): 485 -494

[13] 陈水利. 模糊集理论及其应用[M]. 北京: 科学出版社, 2006: 10-123

[14] 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, 2001: 168-178

[15] Chen Sheng, Cowan C F N, Grant P M. Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks [J]. IEEE Transactions on Neural Networks, 1991, 2: 302-309