不完备信息系统中基于集对相似度的粗集模型

陈圣兵[1] 李龙澍[1,2] 纪 霞[1] 卞世晖[1]
(安徽大学计算智能与信号处理教育部重点实验室 合肥 230039)[1]
(安徽大学计算机科学与技术学院 合肥 230039)[2]

摘 要 讨论了已有粗集扩充模型处理不完备信息的局限,分析了空值相等与确定值相等在概率上的明显差异。依据集对分析理论,提出了集对相似度和相似度容差关系,进而给出一种基于集对相似度的粗集拓展模型。该模型的方法是:通过引入差异度系数体现空值相等与确定值相等之间的差别,利用相似度容差关系及差异度系数确定数据对象的邻域,再利用该邻域得到上下近似集,同时在求上近似时忽略空值的差异性,在求下近似时强调空值的差异性。实验表明,该模型在相同阈值参数的情况下,结果更加合理,精度更高。

关键词 粗集,空值,等价关系,集对分析,不完备信息系统
中图法分类号 TP18 **文献标识码** A

Extension of Rough Set Model Based on SPA Similarity Degree in Incomplete Information Systems

CHEN Sheng-bing[1] LI Long-shu[1,2] JI Xia[1] BIAN Shi-hui[1]
(Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230039, China)[1]
(School of Computer Science and Technology, Anhui University, Hefei 230039, China)[2]

Abstract In view of the limitations of existing extension of rough set models for processing incomplete information, the difference between null value's equality and known value's equality was analysed on probability. Based on the theory of Set Pair Analysis, SPA Similarity Degree and Similarity Tolerance Relation were proposed, and the method of extension of rough set model based on SPA Similarity Degree was described also. It discriminates null value's equality from known value's equality by using the coefficient of difference degree, and gets the neighborhood of object according to coefficient of difference degree and similarity tolerance relation, than gets upper approximation and lower approximation according to the neighborhood. The difference of null value's equality is ignored when we compute upper approximation, and the difference of null value's equality is emphasized for lower approximation. The results of the experiment reveal that both the classification capability and the precision of rough set are better than other models.

Keywords Rough set, Null value, Equivalence relation, Set pair analysis, Incomplete information system

1 引言

基于等价关系的粗集理论(Rough Set Theory, RST)已取得了广泛的应用与发展[9,10],但是这种经典粗集理论的处理对象必须是完备信息系统[1]。而现实的信息系统中,由于各种原因,大量的信息是不完备的,也即存在对象的属性值为空的情况。为此,很多学者对不完备信息系统(Incomplete Information System, IIS)的粗集扩充进行了深入研究。通过削弱等价关系的条件,建立IIS中对象的分类关系,从而将粗集理论应用于IIS,主要涉及基于容差关系[2]、相似关系[3]、限制容差关系[11]等几种扩充模型,但它们都存在一些局限:①对象之间只要有一个已知属性值不同,就被划分在不同的类中,导致知识库中的划分太细,特别不适用于有噪音大型信息系统的处理;②没有体现空值相等的概率,而是假设空值与任意值以概率1相等,造成两个空值较多的对象,其相似可能性很小,但被划分为同一类,这会降低知识库的可信度,尤其不适合精度要求较高的信息系统。

近几年来,随着集对分析(Set Pair Analysis, SPA)在不确定对象上的深入研究[12],出现了利用SPA解决IIS中对象分类的方法。通过集对联系度建立IIS中对象的分类关系,进而求出相应的邻域范围,确定粗集的上、下近似集合。目前,基于SPA的分类关系主要有两种:集对势容差关系和联系度容差关系。集对势容差关系规定差异度为0,仍然没有解决噪音数据、划分过细等问题[13,14];联系度容差关系考察同一度与差异度之和,并用其刻画两个比较对象的联系度[15-17],但这种方法同样会将空值较多的数据对象错误地划分为一类。为此,我们在分析空值相等概率的基础上,提出了一种基于SPA相似度的集对度量方法,并将其应用于IIS。

本文的主要工作从以下几方面展开:(1)通过对各种空值情况的分析,求出空值相等的概率,进而说明空值相等与确定

到稿日期:2009-08-06 返修日期:2009-10-26 本文受国家自然科学基金(60273043),安徽省自然科学基金项目(090412054)资助。
陈圣兵(1973-),男,博士生,讲师,主要研究方向为智能信息处理;李龙澍 教授,博士生导师,主要研究方向为智能软件、知识工程;纪 霞 博士生,主要研究方向为不精确信息处理;卞世晖 硕士生,主要研究方向为人工智能。

值相等的区别;(2)在充分研究集对分析理论的基础上,利用向量空间模型,提出集对相似度的概念,用集对分析方法描述对象的相似性;(3)将该相似度应用于 IIS,由此提出基于集对相似度的粗集扩充模型;(4)利用几组常用的典型数据,将本方法与其他已有方法进行对比,说明本方法对一些特殊对象的划分更加合理;(5)利用 UCI 数据集进行测试,从几个方面考察本方法的整体性能。

2 基于集对理论的分类关系

集对分析由我国学者赵克勤于 20 世纪 80 年代提出[19],其核心思想是从同(同一性)、异(差异不确定性)、反(对立性)3 个侧面研究系统中两个集合在给定问题背景下的确定性联系与不确定性联系的联系、可变与转化,并用一个能充分反映上述情况的同异反联系数系统地开展具体的研究。

2.1 经典 SPA 理论

定义 1 给定两个集合 A 和 B,并设这两个集合组成集对 H = (A,B),在某个具体的问题背景(W)下,H 有 N 个特性,其中 S 个为 A 和 B 所共有,P 个为 A 和 B 相对立,F 个为 A 和 B 既不共有也不对立。集对联系度定义如下:

$$u_W = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j \tag{1}$$

式中,u_W 称为 A 和 B 的联系度,简记 u = a + bi + cj。其中,i 为差异度系数,在[0,1]区间视不同情况取值(有时 i 仅起标记的作用);j 为对立度系数,规定其值恒取 - 1(有时也仅起标记的作用)。一般情况下,联系度 u_W 用式(1)右边的式子表示,只有特殊需要,才对联系度取值(称为联系数)[12]。

2.2 IIS 中基于 SPA 的分类关系

对于 IIS,S = (U,A,V,f),B ⊆ A,设:|B| = n, ∀x,y ∈ U,令 S = {b ∈B | f(x,b) = f(y,b) ≠ *}, F = {b ∈B | f(x,b) = * ∨f(y,b) = *}, P = {b ∈B | (f(x,b) ≠f(y,b)) ∧f(x,b) ≠ * ∧f(y,b) ≠ *},记:

$$u_B = a + bi + cj \tag{2}$$

式中,|B| 表示集合 B 元素的个数,a = |S|/|B|, b = |F|/|B|, c = |P|/|B|, a + b + c = 1。

IIS 中,需要用联系度 u_B 建立对象的分类关系。人们从不同的应用背景提出了两种不同的联系数刻画方法:黄兵、徐怡等认为同类对象不应该有明确不相同的属性值,也即 c = 0, a + b = 1,由此可得 b = 1 - a,进而提出了式(3)的分类关系[13,14];王丽娟、张海东等认为同类对象可以包含少量明确不相同的属性值(噪音数据),亦即 c ≥0, a + b + c = 1,提出了式(4)的分类关系[16,17]:

$$SIM: u = a + bi + cj, c = 0, a \geq \alpha \tag{3}$$

$$SIM: u = a + bi + cj, a + b \geq \alpha \tag{4}$$

上式定义的分类关系有以下几个方面的不足:其一,式(3)不能处理噪音数据,会导致划分过细;其二,式(4)没有充分体现空值相等的概率,而是假定空值和任意值都以概率 1 相等,由此可能将空值较多的数据对象错误地分为一类;其三,它们都只用了集对参数(同一度、差异度和对立度)中的部分数据,从直观上公式缺乏一定的严谨性。

下面先分析 IIS 中空值相等的概率。两个对象 a,b 比较时,在某一属性 A_i 下的属性值 v_i 出现空值的情况有两种:只有一个对象的 v_i 为空值和两个对象的 v_i 都为空值。对于第一种情况,两个对象的 v_i 一个为空值另一个不为空值时,易知其相等的概率为 1/|V_i|,其中 V_i 为属性 A_i 的值域,|V_i| 表示在属性 A_i 上值域元素的个数;当两个对象的 V_i 都为空值时,赵明清[20]等认为其概率为 1/|V_i|^2。但我们发现,两个对象的 V_i 都为空值时,在属性 A_i 上相等的概率为 1/|V_i|,证明如下:

设属性 A_i 下的属性值分别为:$w_1, w_2, \cdots, w_{|V_i|}$,在属性 A_i 上相等有以下|V_i|种可能:

$(v_i)_a = (v_i)_b = w_1$,概率为(1/|V_i|) ×(1/|V_i|) = 1/|V_i|^2

同理,$(v_i)_a = (v_i)_b = w_k$ (k = 2, …,|V_i|)的概率均为 1/|V_i|^2

故两个对象的 v_i 都为空值时,其相等概率为:

$$P = 1/|V_i|^2 + 1/|V_i|^2 + \cdots + 1/|V_i|2 = |V_i| \times (1/|V_i|^2) = 1/|V_i|$$

由以上分析可得如下定理。

定理 1 不完备信息系统 S = (U,A,V,f)中,两个对象的某一个属性 A_i 下的属性值出现空值时,该属性值相等的概率为值域元素数的倒数,记为 1/|V_i|。

由定理 1 可以看出,当两个比较的对象共有 m 个属性出现空值时,其相等的概率为(1/|V_i|)^m。在|V_i|和 m 都较大时,其相似概率值非常小,采用式(3)、式(4)定义的分类关系,可能将它们错误地划分为同一类。为此,我们给出一种新的相似度刻画方法,以根据需要区别差异度和同一度。

2.3 基于集对相似度的容差关系

为了体现空值相等的概率很小的事实,在定义分类关系时,需要通过差异度系数调节空值的可分辨性。同时,考虑到对立度系数 j 值取 - 1 时,同一度与对立度相互抵消对结果会造成影响,提出了集对相似度与相异度。

首先,将式(2)中的集对联系度用向量表示:

$$\overrightarrow{u_B} = (a, bi, c)(x, y, z)^T = (a_x, bi_y, c_z)^T \tag{5}$$

式中,a,b,c 分别为同一度、差异度和对立度;i 为差异系数,可根据情况在[0,1]区间内取值(例如在精度要求很高的背景下取值偏向于 1,而在精度要求很低的背景下取值偏向于 0)。

当构成集对的两个集合完全相同时,$\overrightarrow{u_B} = (1_x, 0_y, 0_z)^T$;当构成集对的两个集合完全不同时,$\overrightarrow{u_B} = (0_x, 0_y, 1_z)^T$。下面,根据向量相似性比较的余弦夹角公式,从相似性和相异性两个方面描述集对属性,分别称为相似度和相异度,定义如下。

定义 2 给定两个集合 A 和 B,a,b,c 分别为其同一度、差异度和对立度,联系度 $u_B = a + bi + cj$ 的向量模型为$\overrightarrow{u_B} = (a_x, bi_y, c_z)^T$,向量$\overrightarrow{u_B}$与 x,y,z 轴的夹角分别为 α, β, γ,其相似度 u_S为$\overrightarrow{u_B}$与向量$(1_x, 0_y, 0_z)^T$夹角的余弦,亦即向量$\overrightarrow{u_B}$在 x 轴方向上的方向余弦 cos α;其相异度 u_D为$\overrightarrow{u_B}$与向量 $(0_x, 0_y, 1_z)^T$ 夹角的余弦,亦即向量$\overrightarrow{u_B}$在 z 轴方向上的方向余弦 cos γ:

$$u_S = \cos\alpha = \frac{a}{\sqrt{a^2 + (bi)^2 + c^2}} \tag{6}$$

$$u_D = \cos\gamma = \frac{c}{\sqrt{a^2 + (bi)^2 + c^2}} \tag{7}$$

由式(6)、式(7)可以看出,随着同一度 a 的增大,差异度 b 和对立度 c 相应减小,相似度 u_S也就相应增大,而相异度 u_D相应减小;反之亦然。

利用上述的集对相似度,定义一个新的二元关系,称为相似度容差关系。

定义3 不完备信息系统 $S=(U,A,V,f)$ 中,设 $B\subseteq A$, $\forall x,y\in U$, $\alpha\in[0,1]$,对象 x,y 的相似度容差关系定义为:

$$SIM(B)=\{\langle x,y\rangle \mid x,y\in U \text{ 且 } u_S(i,x,y)\geq\alpha\}\cup Ix \quad (8)$$

式中,$SIM(B)$ 为相似度容差关系,$u_S(i,x,y)$ 为对象 x,y 的集对相似度,i 为差异度系数,Ix 为恒等函数。易证,由式(8)定义的相似度容差关系具有自反性和对称性,但不具备传递性。

3 基于集对相似度的粗集扩充模型

根据 Pawlak 的粗集理论,知识可理解为集合的划分,由下近似、上近似描述。下近似由确定性成员组成,上近似由可能性成员组成[3]。考虑到 IIS 中空值相等的概率极小,在定义下近似的时候,强调空值相等的差异性,而在定义上近似的时候,忽略空值相等的差异性。

在定义2中,决定相似度 u_S 取值的因素除了 a,b,c 外,还有差异度系数 i,其取值区间为 $[0,1]$。根据以上分析,在基于集对相似度的粗集扩充模型中,求下近似时 i 取最大值1,求上近似时 i 取最小值0。

定义4 不完备信息系统 $S=(U,A,V,f)$ 中,$B\subseteq A$, $\forall x\in U$, $\alpha\in[0,1]$,对象 x 的 i-α 邻域 $S_a^i(x)$ 定义为:

$$S_a^i(x)=\{y\in U\mid u_S(i,x,y)\geq a\}\cup\{x\} \quad (9)$$

式中,i 为差异度系数,取值区间为 $[0,1]$,当 $i=1$ 时,称对象 x 的 i-α 邻域为强 α 邻域,记为 $S_a^1(x)$;当 $i=0$ 时,称对象 x 的 i-α 邻域为弱 α 邻域,记为 $S_a^0(x)$。

由式(6)可得,当 $0\leq i_1\leq i_2\leq 1$ 时,$u_S(i_1,x,y)\geq u_S(i_2,x,y)$,于是,$S_a^{i_1}(x)\supseteq S_a^{i_2}(x)$,$|S_a^{i_1}(x)|\geq|S_a^{i_2}(x)|$。由此可得:

定理2 IIS 中,随着差异度系数 i 的增大,对象 x 的 i-α 邻域 $S_a^i(x)$ 随之减小。也即对于差异度系数 $0\leq i_1\leq i_2\leq 1$,有:

$$S_a^0(x)\supseteq S_a^{i_1}(x)\supseteq S_a^{i_2}(x)\supseteq S_a^1(x)$$

当差异度为0时,也即在完备信息系统中,$S_a^0(x)=S_a^{i_1}(x)=S_a^{i_2}(x)=S_a^1(x)$。

定理3 IIS 中,$S_a^i(x)$ 为对象 x 的 i-α 邻域,$\forall X\subseteq U$,随着差异度系数 i 的增大,$S_a^i(x)$ 包含于集合 X 的概率也随之增大。

证明:设 $0\leq i_1\leq i_2\leq 1$,由定理2可得:

$$S_a^0(x)\supseteq S_a^{i_1}(x)\supseteq S_a^{i_2}(x)\supseteq S_a^1(x)$$

令 P_i 为 $S_a^i(x)\subseteq X$ 的概率,则有:

$P_0\leq P_{i_1}\leq P_{i_2}\leq P_1$,得证。

下面给出 IIS 中基于相似度容差关系的扩充粗集模型。

定义5 不完备信息系统 $S=(U,A,V,f)$ 中,$X\subseteq U$, $B\subseteq A$, $\alpha\in[0,1]$,在 i-α 容差关系上,X 的下近似、上近似分别定义为:

$$\underline{R}_B^a(X)=\{x\mid S_a^1(x)\subseteq X \text{ and } x\in U\} \quad (10)$$

$$\overline{R}_B^a(X)=\{x\mid S_a^0(x)\cap X\neq\Phi \text{ and } x\in U\} \quad (11)$$

在经典粗集理论中,有很多关于下近似集和上近似集的运算性质。根据以上的定义,可以很容易得出以下基于集对相似度的扩充粗集模型的上、下近似集的几个性质。

性质(1): $\underline{R}_B^a(X)\subseteq\overline{R}_B^a(X)$

性质(2): $\underline{R}_B^a(\Phi)=\overline{R}_B^a(\Phi)=\Phi$, $\underline{R}_B^a(U)=\overline{R}_B^a(U)=U$

性质(3): $\underline{R}_B^a(X\cap Y)=\underline{R}_B^a(X)\cap\underline{R}_B^a(Y)$, $\overline{R}_B^a(X\cup Y)=\overline{R}_B^a(X)\cup\overline{R}_B^a(Y)$

性质(4):设 $0\leq\alpha_1\leq\alpha_2\leq 1$,有:

$\underline{R}_{B^{\alpha_1}}^a(X)\subseteq\underline{R}_{B^{\alpha_2}}^a(X)$, $\overline{R}_{B^{\alpha_1}}^a(X)\supseteq\overline{R}_{B^{\alpha_2}}^a(X)$

性质(5):设 $X\subseteq Y\subseteq U$,有:

$\underline{R}_B^a(X)\subseteq\underline{R}_B^a(Y)$, $\overline{R}_B^a(X)\subseteq\overline{R}_B^a(Y)$

性质(6): $\underline{R}_B^a(X\cup Y)\supseteq\underline{R}_B^a(X)\cup\underline{R}_B^a(Y)$, $\overline{R}_B^a(X\cap Y)\subseteq\overline{R}_B^a(X)\cap\overline{R}_B^a(Y)$

在定义5中,对下近似、上近似的定义分别采用了强 α 邻域 $S_a^1(x)$ 和弱 α 邻域 $S_a^0(x)$,亦即差异度系数 i 分别为1和0。由定理3可知,对于论域中任意对象 x,其强 α 邻域属于下近似集的概率最大。于是有如下性质:

性质(7): IIS 中,基于 SPA 相似度容差关系的扩充粗集模型具有最大的下近似集。

性质(7)说明了基于 SPA 相似度的扩充粗集模型的下近似集最大,具有较多的确定性对象,从而保证了较高的精度。

4 实例分析

为了验证本文基于集对相似度的扩充粗集模型的性能,我们将其与当前主流的粗集拓展方法进行比较。分别用一些典型数据和 UCI 公共数据作为处理对象,从不同侧面对这些粗集拓展方法的性能进行分析比较。

4.1 几类典型数据处理能力的比较

首先,选择几类 IIS 中典型的数据,它们在信息处理时极容易出错,成为测量粗集模型优劣的标准之一。根据两个数据对象相等的可能性大小,我们希望实例 Instance1, Instance2 和 Instance3 中的 x,y 两个对象可分辨,实例 Instance4 和 Instance5 中的 x,y 两个对象不可分辨。实例如下:

Instance1: $x_1=(*,*,*,*,*,*,*,h,I,j,k,l,m,n)$,
$y_1=(a,b,c,d,e,f,g,*,*,*,*,*,*,*)$;

Instance2: $x_2=(*,*,*,*,*,*,*,*,*,*,*,*,*,*,*)$,
$y_2=(*,*,*,*,*,*,*,*,*,*,*,*,*,*,*)$;

Instance3: $x_3=(*,*,*,*,*,*,*,*,*,*,*,*,*,*,*)$,
$y_3=(a,b,c,d,e,f,g,h,I,j,k,l,m,n)$;

Instance4: $x_4=(*,b,c,d,e,f,g,h,I,j,k,l,m,n)$,
$y_4=(a,*,c,d,e,f,g,h,I,j,k,l,m,n)$;

Instance5: $x_5=(b,b,c,d,e,f,g,h,I,j,k,l,m,n)$,
$y_5=(a,b,c,d,e,f,g,h,I,j,k,l,m,n)$;

各种粗集扩充模型对于以上数据的处理结果如表1所列。

表1 几种粗集扩充模型的性能比较

扩充模型	Instance1	Instance2	Instance3	Instance4	Instance5
容差关系	不符合	不符合	不符合	符合	不符合
相似关系	符合	不符合	不符合	不符合	不符合
限制容差关系	符合	不符合	符合	符合	不符合
集对势容差关系	符合	符合	符合	符合	不符合
联系度容差关系	不符合	不符合	不符合	符合	符合
相似度容差关系	符合	符合	符合	符合	符合

4.2 分类性能试验结果分析

分类性能是考察一个粗集拓展模型的主要指标。我们利用 UCI 数据集进行测试,比较各种模型在不同程度数据缺失情况下的分类能力。具体步骤为:(1)对于某个完备的数据集 A(数据缺失程度为 0),随机遗失一些记录的属性值(产生存在型空值),得到不完备数据集 A′;(2)对于完备的数据集 A 的某个子集 X,利用等价关系进行划分,得到等价类 R(X);(3)对于不完备数据集 A 的相应的子集 X′,利用容差关系进行划分,得到容差类 T(X′);(4)将集合 R(X)与集合 T(X′)进行相似性比较,相似性越高,说明该模型的整体分类性能越好。比较函数定义如下[9]:

$$u = \frac{|R(X) \cap T(X')|}{|R(X)| + |T(X')| - |R(X) \cap T(X')|} \quad (12)$$

式中,| |表示集合的基数。u ∈[0,1],当 A 和 B 完全相等时,u = 1;完全不同时,u = 0。

实验数据集信息如表 2 所列。

表 2 实验数据集

数据集	样本数	属性个数
Ozone Level Detection	2536	73
Hill-Valley Dataset	606	101

对上述的完备数据集进行离散化处理后,将每个数据集的对象属性值用随机函数定量地遗失 10 %,20 %,30 %,得到空值程度不同的 6 组不完备数据,分别用 OL_10,OL_20,OL_30,HV_10,HV_20,HV_30 表示,利用随机函数产生一定的噪声(实验采用的噪声百分比为 0.4 %),得到相应的不完备数据集。在此数据集上进行各种粗集扩充模型的分类测试,结果如表 3 所列。

表 3 不同空值比例下各种粗集扩充模型的分类能力比较

	OL_10	OL_20	OL_30	HV_10	HV_20	HV_30
容差关系	69.28	65.44	44.52	71.12	53.63	45.50
相似关系	4.15	4.15	4.15	3.99	3.99	3.99
限制容差关系	69.30	65.53	44.52	72.37	53.63	45.50
集对势容差关系	69.43	65.72	44.52	71.12	53.63	33.39
联系度容差关系	76.59	65.13	49.04	78.81	55.34	53.51
相似度容差关系	80.67	72.75	67.59	79.35	59.55	58.16

为了考察各种模型对于不同噪声比数据的分类能力,将[0,0.1]范围 20 等分作为噪声比,利用随机函数,对表 2 所列数据定量加入不同程度的噪声,得到 20 组噪声比不同的不完备数据集合,然后用各种粗集模型进行分类,其分类结果如图 1 所示。

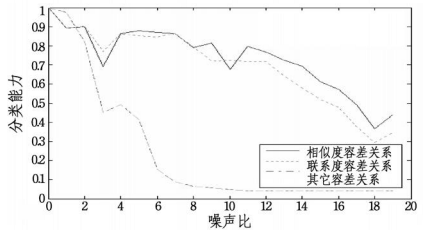

图 1 不同噪声比情况下的分类性能比较

由表 3 可以看出,在不同空值比例的数据集上,基于集对相似度容差关系都得到了较好的分类结果。通过图 1 可以看出,随着数据集噪声比的增大,传统粗集模型的分类能力几乎以同样的速度迅速下降,基于集对理论的联系度容差关系和本文的相似度容差关系分类结果较好,表现出较强的抗噪能力。以下的粗集精度实验将进一步比较联系度模型和相似度模型的性能。

4.3 粗集精度试验分析

基于相似度容差关系的扩充粗集模型的性质(7)说明了该模型下近似集最大,具有较高的精度。为了验证其正确性,对表 2 所列 Hill-Valley Datase 数据用联系度容差关系和相似度容差关系分别进行下近似集求解,并将其精度进行对比。

从数据集 U 中任意取一个子集 X。首先,对集合 X 中的每个对象 x_i,根据定义 4 求对象 x_i 的强 α 邻域 $S_a^1(x_i)$,根据定义 5 求得集合 X 基于相似度容差关系的下近似集 $\underline{R}_B^a(X)$。粗集精度 $a_R(X)$ 定义如下[21]:

$$a_R(X) = \frac{|\underline{R}X|}{|\overline{R}X|} \quad (13)$$

将其精度 $a_R(X)$ 与文献[8]中基于联系度容差关系的下近似集 $\underline{X}_B^a$ 的精度进行比较,结果如图 2 所示。

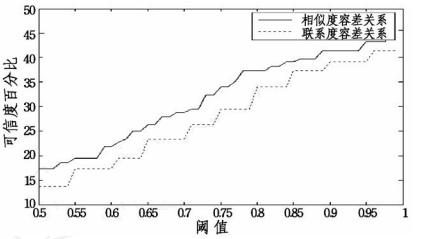

图 2 基于相似度与联系度的粗集精度比较

由图 2 可以看出,基于相似度的粗集模型相对于基于联系度的粗集模型,其粗集的精度较高,这从实验上验证了性质(7)的正确性。

结束语 粗集理论在不完备信息系统应用方面受到的限制,可以通过消弱等价关系的条件使其适用于不完备信息系统。本文在分析已有不完备信息系统的粗集扩充模型(容差关系、相似关系、限制容差关系、集对势容差关系、联系度容差关系)基础之上,利用集对分析理论,提出了集对相似度、相异度测量方法,进而提出了基于相似度容差关系以及对应的粗集扩充模型。一系列的实验结果说明该模型分类能力更强,符合更高的精度要求。本文提出的相似度容差关系在分类过程中,通过引入差异度系数和阈值考虑了根据人的主观要求,同时在求上、下近似时将差异度系数分别取值 0 和 1,使得粗集更加合理,精度更高,这与人机结合以人为主的系统方法论是一致的。

参 考 文 献

[1] Pawlak Z, Grzymala-Busse J, Slowinski R, et al. Rough Sets [J]. Communications of the ACM, 1995, 38(11): 89-95

[2] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1-4): 39-49

[3] Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information[C] //Proc. of the 7th Int'l Workshop on New Directions in Rough Sets, Data Mining, and Granular Soft Computing. Berlin: Springer-Verlag, 1999: 73-81

[4] Yang Xibei, Yang Jingyu, Wu Chen, et al. Dominance-based rough set approach and knowledge reductions in incomplete ordered information system[J]. Information Sciences: an International Journal, 2008, 178(4): 1219-1234

[5] Thangavel K, Pethalakshmi A. Dimensionality reduction based

on rough set theory: A review[J]. Applied Soft Computing, 2009,9(1):1568-4946

[6] Leung Yee, Fung Tung, Mi Ju-Sheng, et al. A rough set approach to the discovery of classification rules in spatial data [J]. International Journal of Geographical Information Science,2007,21(9):1033-1058

[7] Yang Xibei, Qu Fang, Yang Jingyu, et al. A Novel Extension of Rough Set Model in Incomplete Information System[C]//Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control. June 2008:306-312

[8] Sabu M K, Raju G. Rough Set Approaches for Mining Incomplete Information Systems[C]//Proceedings of the 4th International Conference on Intelligent Computing. Sep. 2008:914-921

[9] 张文明,薛青. 粗糙集方法在作战仿真数据挖掘中的应用[J]. 系统仿真学报,2006,18(2):179-181

[10] 于艾清,顾幸生. 基于广义粗糙集的不确定条件下的 Flow Shop 调度[J]. 系统仿真学报,2006,18(12):3369-3376

[11] 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展,2002,39(10):1238-1243

[12] 赵克勤. 集对分析及其初步应用[M]. 杭州:浙江科学出版社,2000

[13] 黄兵,周献中. 不完备信息系统中基于联系度的粗集模型拓展[J]. 系统工程理论与实践,2004(1):88-92

[14] 徐怡,李龙澍,李学俊. 基于集对势的扩充粗糙集模型[J]. 系统仿真学报,2008,20(6):97-103

[15] 黄兵,周献中. 基于集对分析的不完备信息系统粗糙集模型[J]. 计算机科学,2002,29(9 专刊):1-3

[16] 王丽娟,吴陈,严熙. 基于限制容差关系和集对分析的数据依赖在 IIS 中的应用[J]. 系统工程理论与实践,2007,11:97-103

[17] 张海东,舒兰. 限制容差关系下的集对变精度粗糙集模型[J]. 模糊系统与数学,2007,21(5):125-130

[18] 杨习贝,杨静宇,於东军,等. 不完备信息系统中的可变精度分类粗糙集模型[J]. 系统工程理论与实践,2008,28(5):116-121

[19] 赵克勤. SPA 的同异反系统理论在人工智能研究中的应用[J]. 智能系统学报,2007,2 (5):20-35

[20] 赵明清,胡美燕,郭世伟. 量化容差关系与量化非对称相似关系的比较研究[J]. 计算机科学,2004,30(10):98-100

[21] 张文修. 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 124 页)

网络中各个节点的实际最大流如表 3 所列。

表 3 网络中各个节点的实际最大流

节点	1	2	3	4	5	6	8	9
最大流	3	2	7	6	7	7	4	5
节点	10	12	14	15	16	17	18	19
最大流	5	7	7	3	7	2	6	6

模拟实验 2 的结果如图 6 所示。

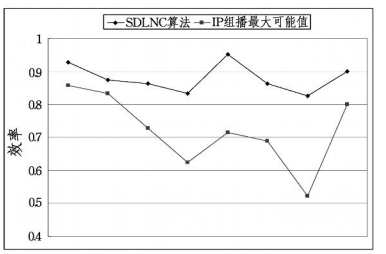

图 6 模拟实验 2 的运行结果

算法的效率表示 SDLNC 算法的组播容量和网络拓扑理论上最大的容量的比值。从图 4 - 图 6 的普适性实验结果可以看出,无论 R 如何选择,效率均在 0.8 以上,比同样条件下的 IP 组播要高。汇点的差异越大,SDLNC 算法的优势就越明显。这充分说明了 SDLNC 算法的高效率并不依赖于 R 的选择,而是网络编码算法的优越性带来的提高。

结束语 本文基于现有的网络编码算法,实现了一种组播网络的路由算法——SDLNC 算法。SDLNC 算法利用分层编码理论,充分满足了异质性节点对带宽的不同需求,并初步解决了同类算法中存在的一些问题。它的特点是网络中节点不需要知道周围节点的情况,能在分布式计算的过程中逐步获取编码及路由所需要的信息。仿真试验的结果表明,本算法在不同的网络拓扑下都有不错的性能,均接近或达到了理论上限值。

但是 SDLNC 算法也存在着一些不足之处,并没有完全实现数学模型中的目标,还有很大的优化和改进的空间。

参 考 文 献

[1] Sarrafzadeh M, Wong C K. Bottleneck Steiner Trees in the Plane [J]. IEEE Transactions on Computers,1992,41(3):370-374

[2] RFC2189. Core Based Trees (CBT version 2) Multicast Routing[S]

[3] Ahlswede R, Cai Ning, Li S-Y R, et al, Network Information Flow [J]. IEEE Transactions on Information Theory,2000,46(4):1204-1216

[4] Sanders P, Egner S, Tolhuizen L. Polynomial Time Algorithms for Network Information Flow [C]//Proceedings of the 15th annual ACM Symposium on Parallel Algorithms and Architectures. San Diego California USA,2003:286-294

[5] Sundaram N, Ramanathan P, Banerjee S. Multirate Media Streaming Using Network Coding [C]//Proc. 43rd Allerton Conference on Communication Control and Computing. Monticello IL,Sep. 2005

[6] Koetter R, Medard M. An Algebraic Approach to Network Coding [J]. IEEE/ACM Transactions on Networking, 2003, 11(5):782-795