基于特征矩阵的粗糙集关系性质分析与表示定理*)

张晓如　张再跃

(江苏科技大学电子信息学院　江苏镇江 212003)

摘　要　本文着重研究粗糙集理论基本概念与基本运算的矩阵表示，用特征矩阵描述粗糙集理论中的基本概念，并通过研究特征矩阵运算性质，揭示和刻画粗糙集知识空间的基本代数性质。同时，定义特征矩阵“与积”和“或积”两种逻辑运算，分别对上、下近似概念相对应；针对完备信息系统与不完备信息系统的特点，分析基于对象属性值的相关关系性质，证明不同关系下对象集的上、下近似集和对象关系类的特征矩阵表示定理。

关键词　粗糙集，特征矩阵，逻辑运算，信息系统

Analysis of Relational Properties and Representing Theorems of Rough Set Based on Characteristic Matrix

ZHANG Xiao-Ru　ZHANG Zai-Yue

(School of Electronics and Information of Jiangsu University of Sci. &Tech., Zhenjiang, Jiangsu212003)

Abstract　In this paper, the basic notations in the Rough Sets are discussed by using the matrix method. In order to give out a complete matrix description of rough set and its calculations, the both Boolean calculations “and-product” and “or-product”, corresponding separately to the notations of the up-approximation and the lower-approximation, between the characteristic matrixes are defined. Aim at the characteristics of the complete and non-complete information systems, the associated relational properties are discussed, and the characteristic matrix representing theorems about the up-lower-approximation of rough set and the relational class of the object are proved.

Keywords　Rough set, Characteristic matrix, Boolean calculation, Information system

1　引言

粗糙集(Rough Sets)理论是由波兰学者 Pawlak 教授在20世纪80年代提出的研究不完整、不确定知识和数据的表达、学习、归纳的理论方法[1]，现已成为知识工程研究领域一种有效的数学工具，并在模式识别、机器学习、决策支持、过程控制、预测建模等许多科学与工程领域得到成功的应用[2]。

矩阵理论作为一种基本的数学工具，在数值分析、优化理论、微分方程、概率统计、控制论、网络等学科领域有着十分重要的应用。随着计算机科学与计算技术的发展，知识工程领域相关问题的研究备受关注，矩阵理论与方法在知识表示与知识推理等方面得到广泛应用。在基于粗糙集理论的知识表示与知识获取研究方面，人们从各种侧面，对矩阵的理论与方法加以运用并取得成果。如通过矩阵运算对基于等价关系的知识基中对象集的上、下近似算子进行表述，分析和研究粗糙集的基本性质[3]；运用矩阵方法描述信息系统属性间的依赖关系[4]，分析和研究不同情况下知识属性的约简方法等等[5~7]。

本文着重研究粗糙集理论基本概念与基本运算的矩阵表示，用特征矩阵对粗糙集理论中的基本概念进行描述，并通过特征矩阵运算性质研究，揭示和刻画粗糙集知识空间的基本代数性质。第2节，简要给出了粗糙集的基本概念与基本代数性质；第3节，定义了包括“与积”和“或积”在内的特征矩阵的逻辑运算，分析了运算的基本性质；第4、第5节，分别针对完备信息系统与不完备信息系统的特点，分析了基于对象属性值的多种关系的性质，给出了不同关系下对象集的上、下近似集和对象关系类的特征矩阵表示定理；最后为小结。

2　粗糙集基本概念与基本代数性质

粗糙集理论研究的基本内容是基于属性特征的知识与信息系统，其代数结构的基本表示形式是一知识基 $K=\langle U,R\rangle$，其中 U 为论域，R 为 U 上的关系。它的最基本情形是依据对象的属性值对论域 U 的一种划分，称为 U 上的不分明关系，是粗糙集理论的一个关键概念。粗糙集理论的精华部分是利用 R 定义了 $P(U)$ 上上近似算子 R^* 和下近似算子 R_* 的概念，即对任意集合 $X\in P(U)$，

$R_*(X)=\{x|(x\in U\wedge[x]_R\subseteq X)\}$

$R^*(X)=\{x|(x\in U\wedge[x]_R\cap X\neq\phi)\}$

其中 $[x]_R=\{y|(x,y)\in R\}$，是 x 关于 R 所在的等价类。如下命题给出粗糙集的基本代数性质：

命题 2.1

(1)$R_*(X)\subseteq X\subseteq R^*(X)$

(2)$R_*(\phi)=R^*(\phi)=\phi, R_*(U)=R^*(U)=U$

(3)$R^*(X\cup Y)=R^*(X)\cup R^*(Y)$

(4)$R_*(X\cap Y)=R_*(X)\cap R_*(Y)$

(5)$X\subseteq Y\Rightarrow R_*(X)\subseteq R_*(Y)$

(6)$X\subseteq Y\Rightarrow R^*(X)\subseteq R^*(Y)$

(7)$R_*(X\cup Y)\supseteq R_*(x)\cup R_*(Y)$

(8)$R^*(X\cap Y)\subseteq R^*(x)\cap R^*(Y)$

(9)$R_*(\overline{X})=\overline{R^*(X)}$

*)国家自然科学基金项目(No. 60543064)资助。**张晓如**　副教授，主要研究方向：智能信息处理、基础数学；**张再跃**　教授，主要研究方向：理论计算机科学、智能信息处理。

(10)$R^*(\overline{X})=\overline{R_*(X)}$

(11)$R_*(R_*(X))=R^*(R_*(X))=R_*(X)$

(12)$R^*(R^*(X))=R_*(R^*(X))=R^*(X)$

3 特征矩阵的逻辑运算与基本性质

与($\wedge$)、或($\vee$)、非($-$)是基本的逻辑运算。在二值逻辑中,对任意$a,b\in\{0,1\}$有$a\wedge b=1$当且仅当$a=1$并且$b=1$;$a\vee b=0$当且仅当$a=0$并且$b=0$;$\bar{a}=1$当且仅当$a=0$。本节将以此为基础,建立特征矩阵间的逻辑运算,并对运算的性质进行分析。

定义 3.1 设$A=(a_{ij})_{m\times n}$是一$m\times n$矩阵。如果对任意i和j,都有$a_{ij}\in\{0,1\}$,则称A是特征矩阵。

定义 3.2 设$A=(a_{ij})$和$B=(b_{ij})$是两个$m\times n$特征矩阵,则$m\times n$特征矩阵$C=(a_{ij}\wedge b_{ij})$称为A和B的交,记为$C=A\wedge B$。

定义 3.3 设$A=(a_{ij})$和$B=(b_{ij})$是两个$m\times n$特征矩阵,则$m\times n$特征矩阵$C=(a_{ij}\vee b_{ij})$称为A和B的并,记为$C=A\vee B$。

定义 3.4 设$A=(a_{ij})$是$m\times n$特征矩阵,则$m\times n$特征矩阵$C=(\bar{a}_{ij})$称为A的非,记为$C=\bar{A}$。

性质 3.1 对任意$m\times n$特征矩阵$A=(a_{ij})$,$B=(b_{ij})$和$C=(c_{ij})$,有如下性质成立:

(1)$A\wedge A=A$,$A\vee A=A$;

(2)$A\wedge B=B\wedge A$,$A\vee B=B\vee A$;

(3)$A\wedge(B\vee C)=(A\wedge B)\vee(A\wedge C)$,$A\vee(B\wedge C)=(A\vee B)\wedge(A\vee C)$;

(4)$\overline{A\wedge B}=\bar{A}\vee\bar{B}$,$\overline{A\vee B}=\bar{A}\wedge\bar{B}$。

定义 3.5 设$A=(a_{ij})$和$B=(b_{ij})$分别是两个$m\times p$和$p\times n$特征矩阵。如果$m\times n$特征矩阵$C=(c_{ij})$满足$c_{ij}=(a_{i1}\wedge b_{1j})\vee\cdots\vee(a_{ip}\wedge b_{pj})$(记为$\sum_{k=1}^{p}(a_{ik}\wedge b_{kj})$),则称$C$为$A$和$B$的与积,记为$C=A\otimes B$。

定义 3.6 设$A=(a_{ij})$和$B=(b_{ij})$分别是两个$m\times p$和$p\times n$特征矩阵。如果$m\times n$特征矩阵$C=(c_{ij})$满足$c_{ij}=(a_{i1}\vee b_{1j})\wedge\cdots\wedge(a_{ip}\vee b_{pj})$(记为$\prod_{k=1}^{p}(a_{ik}\vee b_{kj})$),则称$C$为$A$和$B$的或积,记为$C=A\oplus B$。

性质 3.2 对任意$m\times p$特征矩阵$A=(a_{ij})$,$p\times n$特征矩阵$B=(b_{ij})$和$n\times k$特征矩阵$C=(c_{ij})$,有如下性质成立:

(1)$A\otimes(B\otimes C)=(A\otimes B)\otimes C$,$A\oplus(B\oplus C)=(A\otimes B)\oplus C$;

(2)$\overline{A\otimes B}=\bar{A}\oplus\bar{B}$,$\overline{A\oplus B}=\bar{A}\otimes\bar{B}$。

性质 3.3 对任意$m\times p$特征矩阵$A=(a_{ij})$,$p\times n$特征矩阵$B=(b_{ij})$和$C=(c_{ij})$,有如下性质成立:

(1)$A\otimes(B\vee C)=(A\otimes B)\vee(A\otimes C)$;

(2)$A\oplus(B\wedge C)=(A\oplus B)\wedge(A\oplus C)$。

4 特征矩阵运算与粗糙集基本代数性质的刻画

从本节开始我们将用特征矩阵对粗糙集理论中的基本概念进行描述,文中的特征矩阵用粗体黑体英文字母表示。对此,设$K=\langle U,R\rangle$是一知识基,其中$U=\{e_1,\cdots,e_n\}$是知识基K的论域,R是U上的等价关系。对任意$x\in U$,$[x]_R$表示对象x代表的关于R的等价类。

定义 4.1 如果特征矩阵$\boldsymbol{R}=(r_{ij})_{n\times n}$满足$r_{ij}=1$当且仅当$(e_i,e_j)\in R$,则$\boldsymbol{R}$称为$U$上等价关系$R$的特征矩阵。

定义 4.2 设$X\subseteq U$是U的子集,如果$n\times 1$特征矩阵$\boldsymbol{X}=(x_1,\cdots,x_n)'$满足$x_i=1$当且仅当$e_i\in X$,则称$\boldsymbol{X}$是集合$X$的特征矩阵,其中$(x_1,\cdots,x_n)'$表示$(x_1,\cdots,x_n)$的转置。$X$的特征矩阵$\boldsymbol{X}$也称为$X$的特征向量。特别,$n\times 1$特征矩阵$\boldsymbol{Ex}$表示单点集$\{x\}\subseteq U$的特征向量。

在粗糙集理论中,幂集$P(U)$上的上近似算子R^*和下近似算子R_*扮演着极其重要的角色。对任意集合$X\in P(U)$,用$\boldsymbol{R^*X}$和$\boldsymbol{R_*X}$分别表示R^*X和R_*X的特征向量,运用特征矩阵运算我们有如下表示定理。

定理 4.1 (1)$\boldsymbol{R}\otimes\boldsymbol{X}=\boldsymbol{R^*X}$;(2)$\bar{\boldsymbol{R}}\oplus\boldsymbol{X}=\boldsymbol{R_*X}$。

证明:令$\boldsymbol{R}=(r_{ij})$,$X=(x_1,\cdots,x_n)'$。

(1)设$\boldsymbol{R^*X}=(a_1,\cdots,a_n)'$,则对任意$i$,$a_i=1$当且仅当$e_i\in R^*(X)$,当且仅当$[e_i]_R\cap X\neq\phi$,当且仅当$\exists k(e_k\in[e_i]\wedge e_k\in X)$,即$\exists k(r_{ik}=1\wedge x_k=1)$,当且仅当$\sum_{k=1}^{n}(r_{ik}\wedge x_k)=1$。

(2)设$\boldsymbol{R_*X}=(b_1,\cdots,b_n)'$,则对任意$i$,$b_i=1$当且仅当$e_i\in R_*(x)$,当且仅当$[e_i]_R\subseteq X$,当且仅当$\forall k(e_k\in[e_i]\rightarrow e_k\in X)$,即$\forall k(r_{ik}=1\rightarrow x_k=1)$,从而有$\forall k(\bar{r}_{ik}\vee x_k=1)$,当且仅当$\prod_{k=1}^{n}(\bar{r}_{ik}\vee x_k)=1$。 □

在粗糙集理论中,通常需要考虑集合间的包含关系。此外,对象集U上关系R的性质尤为重要,不同性质的关系所产生的对象分类也不同,因而产生不同的知识结构。对此,我们将根据R的特点研究其特征矩阵$\boldsymbol{R}$的性质,当关系R具有某种性质时,如自反性、对称性等,我们也称$\boldsymbol{R}$是具有某种性质的特征矩阵。

性质 4.2 设$\boldsymbol{R}=(x_{ij})$为U上关系R的特征矩阵。如果$\boldsymbol{R}$是自反的,则对任意i有$x_{ii}=1$;如果$\boldsymbol{R}$是对称的,则对任意i,j均有$x_{ij}=x_{ji}$;如果$\boldsymbol{R}$是传递的,则对任意i,j,k均有$\bar{x}_{ik}\vee\bar{x}_{kj}\vee x_{ij}=1$。

定义 4.3 设$\boldsymbol{R}=(x_{ij})$,$\boldsymbol{Q}=(y_{ij})$为两同阶特征矩阵,如果对任意i,j均有$\bar{x}_{ij}\vee y_{ij}=1$,则称矩阵$\boldsymbol{R}$小于等于矩阵$\boldsymbol{Q}$,记为$\boldsymbol{R}\leqslant\boldsymbol{Q}$。

显然,如果$\boldsymbol{R}\leqslant\boldsymbol{Q}$那么$\bar{\boldsymbol{Q}}\leqslant\bar{\boldsymbol{R}}$。根据定义,我们不难验证如下事实:

命题 4.3 设$X,Y\in P(U)$,$\boldsymbol{X}=(x_1,\cdots,x_n)'$,$\boldsymbol{Y}=(y_1,\cdots,y_n)'$分别为它们的特征向量。则$X\subseteq Y$当且仅当$\boldsymbol{X}\leqslant\boldsymbol{Y}$。

命题 4.4 设$\boldsymbol{X}=(x_1,\cdots,x_n)'$,$\boldsymbol{Y}=(y_1,\cdots,y_n)'$为特征向量,$\boldsymbol{R}$为任意$n$阶特征矩阵。若$\boldsymbol{X}\leqslant\boldsymbol{Y}$,则$\boldsymbol{R}\oplus\boldsymbol{X}\leqslant\boldsymbol{R}\oplus\boldsymbol{Y}$,$\boldsymbol{R}\otimes\boldsymbol{X}\leqslant\boldsymbol{R}\otimes\boldsymbol{Y}$。

命题 4.5 设$\boldsymbol{X}=(x_1,\cdots,x_n)'$为特征向量,$\boldsymbol{R}$和$\boldsymbol{Q}$为任意$n$阶特征矩阵。若$\boldsymbol{R}\leqslant\boldsymbol{Q}$,则$\boldsymbol{Q}\oplus\boldsymbol{X}\leqslant\boldsymbol{R}\oplus\boldsymbol{X}$,$\boldsymbol{R}\otimes\boldsymbol{X}\leqslant\boldsymbol{Q}\otimes\boldsymbol{X}$。

命题 4.6 设R是U上的等价关系,$\boldsymbol{R}=(r_{ij})$是R的特征矩阵,则

(1)$\boldsymbol{R}\otimes\boldsymbol{R}=\boldsymbol{R}$,$\bar{\boldsymbol{R}}\oplus\bar{\boldsymbol{R}}=\bar{\boldsymbol{R}}$;

(2)$\boldsymbol{R}\otimes\bar{\boldsymbol{R}}=\bar{\boldsymbol{R}}$,$\bar{\boldsymbol{R}}\oplus\boldsymbol{R}=\boldsymbol{R}$;

在第2节中我们列出了粗糙集的基本代数性质。运用特征矩阵运算的基本性质,对应命题2.1即可得到如下命题:

命题 4.7 设$K=\langle U,R\rangle$是一知识基,其中$U=\{e_1,\cdots,e_n\}$是知识基K的论域,R是U上的等价关系,$\boldsymbol{R}=(r_{ij})$表示R的特征矩阵,$\boldsymbol{U}$表示论域U的特征向量,$\boldsymbol{\Phi}$表示空子集$\phi\subseteq U$的特征向量,$\boldsymbol{X}$表示子集$X\subseteq U$的特征向量。则

(1)$\bar{\boldsymbol{R}}\oplus\boldsymbol{X}\leqslant\boldsymbol{X}\leqslant\boldsymbol{R}\otimes\boldsymbol{X}$;

(2)$\bar{\boldsymbol{R}}\oplus\boldsymbol{\Phi}=\boldsymbol{R}\otimes\boldsymbol{\Phi}=\boldsymbol{\Phi}$,$\bar{\boldsymbol{R}}\oplus\boldsymbol{U}=\boldsymbol{R}\otimes\boldsymbol{U}=\boldsymbol{U}$;

(3)$\boldsymbol{R}\otimes(\boldsymbol{X}\vee\boldsymbol{Y})=(\boldsymbol{R}\otimes\boldsymbol{X})\vee(\boldsymbol{R}\otimes\boldsymbol{Y})$;

(4)$\bar{\boldsymbol{R}}\oplus(\boldsymbol{X}\wedge\boldsymbol{Y})=(\bar{\boldsymbol{R}}\oplus\boldsymbol{X})\wedge(\bar{\boldsymbol{R}}\oplus\boldsymbol{Y})$;

(5)$\boldsymbol{X}\leqslant\boldsymbol{Y}\Rightarrow\bar{\boldsymbol{R}}\oplus\boldsymbol{X}\leqslant\bar{\boldsymbol{R}}\oplus\boldsymbol{Y}$;

(6)$X\leqslant Y\Rightarrow R\otimes X\leqslant R\otimes Y$;

(7)$\bar{R}\oplus X\vee\bar{R}\oplus Y\leqslant\bar{R}\oplus(X\vee Y)$;

(8)$R\otimes(X\wedge Y)\leqslant R\otimes X\wedge R\otimes Y$;

(9)$\bar{R}\oplus\bar{X}=\overline{R\otimes X}$

(10)$R\otimes\bar{X}=\overline{\bar{R}\oplus X}$

(11)$\bar{R}\oplus(\bar{R}\oplus X)=R\otimes(\bar{R}\oplus X)=\bar{R}\oplus X$;

(12)$R\otimes(R\otimes X)=\bar{R}\oplus(R\otimes X)=R\otimes X$。

5 不完备信息系统的特征矩阵表示

随着对不完备信息系统研究的深入,粗糙集理论得到扩展。在不完备信息系统中,由于某些对象的某些属性值难以确定,因而用于对象分类的关系性质也随之发生变化。本节我们将运用特征矩阵对扩展的粗糙集理论中基本概念与运算加以表示。

设 $K=\langle U,C,V,f\rangle$ 是一信息表,其中 $U=\{e_1,\cdots,e_n\}$ 是论域,$C=\{c_1,\cdots,c_m\}$ 是属性集,$V=\bigcup_{c_i\in C}Vc_i$ 是属性值集,$f:U\times C\to V$ 是信息函数。在不完备信息表中,通常用符号 * 表示对象关于某属性的遗漏或无法确定值。

定义 5.1 基于信息表 K 上的容差关系 $T\subseteq U\times U$ 定义为

$$T=\{(x,y)\mid\forall c_i\in C(f(x,c_i)=f(y,c_i)\vee f(x,c_i)=*\vee f(y,c_i)=*)\}$$

且对任意 $x\in U$,$I_C(x)=\{y\in U\mid(x,y)\in T\}$ 表示 x 关于 T 的容差类。

对任意 $X\in P(U)$,其基于容差关系 T 的上、下近似定义为

$$T^*(X)=\{x\mid(x\in U\wedge I_C(x)\cap X\neq\phi)\}$$

$$T_*(X)=\{x\mid(x\in U\wedge I_C(x)_R\subseteq X)\}$$

我们沿用第 4 节的有关符号,同时用 $\boldsymbol{T}$ 表示关系 T 的特征矩阵,$\boldsymbol{T}^*X$,$\boldsymbol{T}_*X$ 和 $\boldsymbol{I}_C(x)$ 分别表示集合 X 的上、下近似以及对象 x 容差类的特征向量,则有如下表示定理成立:

定理 5.1 (1)对任意 $e_i\in U$,$(\boldsymbol{I}_C(e_i))'=(\boldsymbol{E}e_i)'\otimes\boldsymbol{T}$ 或 $\boldsymbol{I}_C(e_i)=\boldsymbol{T}\otimes\boldsymbol{E}e_i$。

(2)对任意 $X\in P(U)$,$\boldsymbol{T}^*X=\boldsymbol{T}\otimes X$ 或 $(\boldsymbol{T}^*X)'=X'\otimes\boldsymbol{T}$;$\boldsymbol{T}_*X=\bar{\boldsymbol{T}}\oplus X$ 或 $(\boldsymbol{T}_*X)'=X'\oplus\bar{\boldsymbol{T}}$。

证明:(1)令 $\boldsymbol{T}=(t_{ij})$,$(\boldsymbol{E}e_i)'=(x_1,\cdots,x_n)$,$(\boldsymbol{I}_c(e_i))'=(a_1,\cdots,a_n)$。则对任意 j,$a_j=1$ 当且仅当 $e_j\in I_C(e_i)$,当且仅当 $(e_i,e_j)\in T$,当且仅当 $t_{ij}=1$。注意到 $x_i=1$,因而有 $x_i\wedge t_{ij}=1$,故 $\sum_{k=1}^n(x_k\wedge t_{kj})=1$;反之,若 $\sum_{k=1}^n(x_k\wedge t_{kj})=1$,由于其中只有 $x_i=1$,因而必有 $t_{ij}=1$,继而有 $a_j=1$,由此证得 $(\boldsymbol{I}_C(e_i))'=(\boldsymbol{E}e_i)'\otimes\boldsymbol{T}$。因为 $\boldsymbol{T}$ 对称,所以又有 $\boldsymbol{I}_C(e_i)=\boldsymbol{T}\otimes\boldsymbol{E}e_i$。

(2)$\boldsymbol{T}^*X=\boldsymbol{T}\otimes X$ 和 $\boldsymbol{T}_*X=\bar{\boldsymbol{T}}\oplus X$ 的证明参见定理 4.1,$(\boldsymbol{T}^*X)'=X'\otimes\boldsymbol{T}$ 和 $(\boldsymbol{T}_*X)'=X'\oplus\bar{\boldsymbol{T}}$ 可利用 $\boldsymbol{T}$ 的对称性得到。□

在不完备信息表处理中,由于容差关系的分类往往显得粗糙,人们又在此基础上引入了非对称相似关系,其定义如下:

定义 5.2 基于信息表 K 上的非对称相似关系 $S\subseteq U\times U$ 定义为

$$S=\{(x,y)\mid\forall c_i\in C(f(x,c_i)=f(y,c_i)\vee f(x,c_i)=*)\}$$

显然 S 是自反和传递的,但不对称。所以在分析两对象之间的关系时,我们必须从两个方面予以考虑。对此,对任意 $x\in U$,我们分别用 $J_C(x)=\{y\in U\mid(y,x)\in S\}$ 和 $J_C^{-1}(x)=\{y\in U\mid(x,y)\in S\}$ 表示非对称相似于 x 和 x 非对称相似的对象集合。且对任意 $X\in P(U)$,其关于非对称相似关系 S 的上、下近似定义为

$$S^*(X)=\bigcup_{x\in X}J_C(x);$$

$$S_*(X)=\{x\mid(x\in U\wedge J_C^{-1}(x)\subseteq X)\}$$

命题 5.2 对任意 $x,y\in U$,$x\in J_C(y)$ 当且仅当 $y\in J_C^{-1}(x)$。

命题 5.3 对任意 $X\subseteq U$,$S^*(X)=\bigcup_{x\in X}J_C(x)=\{y\mid y\in U\wedge J_C^{-1}(y)\cap X\neq\phi\}$。

证明:对任意 $y\in U$,$y\in S^*(X)$ 当且仅当 $\exists x(x\in X\wedge y\in J_C(x))$,当且仅当 $\exists x(x\in X\wedge x\in J_C^{-1}(y))$,当且仅当 $J_C^{-1}(y)\cap X\neq\phi$。□

我们用 $\boldsymbol{S}$ 表示关系 S 的特征矩阵,$\boldsymbol{S}^*X$,$\boldsymbol{S}_*X$,$\boldsymbol{J}_C(X)$,$\boldsymbol{J}_C^{-1}(X)$ 分别表示集合 X 关于非对称相似关系 S 的上、下近似以及非对称相似于 x 和 x 非对称相似的对象集合的特征向量,则有如下表示定理成立:

定理 5.4 对任意 $e_i\in U$,$\boldsymbol{J}_C(e_i)=\boldsymbol{S}\otimes\boldsymbol{E}e_i$;$\boldsymbol{J}_C^{-1}(e_i)=\boldsymbol{S}'\otimes\boldsymbol{E}(e_i)$,其中 $\boldsymbol{S}'$ 为特征矩阵 $\boldsymbol{S}$ 的转置。

证明:(1)设 $\boldsymbol{S}=(s_{ij})$,$\boldsymbol{E}e_i=(x_1,\cdots,x_n)'$,$\boldsymbol{J}_C(e_i)=(a_1,\cdots,a_n)'$。则对任意 j,$a_j=1$ 当且仅当 $e_j\in J_C(e_i)$,当且仅当 $(e_j,e_i)\in S$,当且仅当 $s_{ji}=1$。注意到 $x_i=1$,因而有 $s_{ji}\wedge x_i=1$,故 $\sum_{k=1}^n(s_{jk}\wedge x_k)=1$;反之,若 $\sum_{k=1}^n(s_{jk}\wedge x_k)=1$,由于其中只有 $x_i=1$,因而必有 $s_{ji}=1$,即 $(e_j,e_i)\in S$,所以有 $a_j=1$,由此证得 $\boldsymbol{J}_C(e_i)=\boldsymbol{S}\otimes\boldsymbol{E}e_i$。

又设 $\boldsymbol{S}'=(s'_{ij})$ 为特征矩阵 $\boldsymbol{S}$ 的转置,$\boldsymbol{J}_C^{-1}(e_i)=(b_1,\cdots,b_n)'$。则对任意 j,由于 $x_k=1$ 当且仅当 $k=i$,因此 $\sum_{k=1}^n(s'_{jk}\wedge x_k)=1$ 当且仅当 $s'_{ji}=s_{ij}=1$,即有 $(e_i,e_j)\in S$,$e_j\in J_C^{-1}(e_i)$,从而 $b_j=1$,由此证得 $\boldsymbol{J}_C^{-1}(e_i)=\boldsymbol{S}'\otimes\boldsymbol{E}e_i$。□

定理 5.5 对任意 $X\in P(U)$,$\boldsymbol{S}^*X=\boldsymbol{S}\otimes X$;$\boldsymbol{S}_*X=\bar{\boldsymbol{S}}\oplus X$。

证明:令 $\boldsymbol{S}=(s_{ij})$,$\boldsymbol{X}=(x_1,\cdots,x_n)'$。

设 $\boldsymbol{S}^*X=(a_1,\cdots,a_n)'$,则对任意 i,$a_i=1$ 当且仅当 $e_i\in S^*(X)=\bigcup_{x\in X}J_C(x)$,当且仅当 $\exists j(e_j\in X\wedge e_i\in J_C(e_j))$,即 $\exists j(r_{ij}=1\wedge x_j=1)$,当且仅当 $\sum_{k=1}^n(r_{ik}\wedge x_k)=1$。

若 $\boldsymbol{S}_*X=(b_1,\cdots,b_n)'$,则对任意 i,$b_i=1$ 当且仅当 $e_i\in S_*(X)$,当且仅当 $J_C^{-1}(e_i)\subseteq X$,当且仅当 $\forall k(e_k\in J_C^{-1}(e_i)\to e_k\in X)$,即 $\forall k(r_{ik}=1\to x_k=1)$,从而有 $\forall k(\bar{r}_{ik}\vee x_k=1)$,当且仅当 $\prod_{k=1}^n(\bar{r}_{ik}\vee x_k)=1$。□

在一给定的信息表 $K=\langle U,C,V,f\rangle$ 中,当所考虑的属性集 C 确定时,非对称相似关系 S 所满足的条件是容差关系 T 所满足条件的子集,即对任意 $x,y\in U$,当 $(x,y)\in S$ 时,必有 $(x,y)\in T$,所以若用 $\boldsymbol{S}=(s_{ij})$ 和 $\boldsymbol{T}=(t_{ij})$ 分别表示 S 和 T 的特征矩阵,则有 $s_{ij}=1\to t_{ij}=1$,即 $\bar{S}_{ij}\vee t_{ij}=1$,根据定义 4.3 有 $S\leqslant T$ 和 $\bar{T}\leqslant\bar{S}$。于是对任意子集 $X\subseteq U$ 的特征向量 $\boldsymbol{X}$,根据命题 4.5,就有 $\bar{\boldsymbol{T}}\oplus\boldsymbol{X}\leqslant\bar{\boldsymbol{S}}\oplus\boldsymbol{X}$ 和 $\boldsymbol{S}\otimes\boldsymbol{X}\leqslant\boldsymbol{T}\otimes\boldsymbol{X}$,再根据表示定理 5.1(2)和定理 5.5,即可得到 $T_*X\subseteq S_*X$ 和 $S^*X\subseteq T^*X$。对此,我们可以认为:

推论 5.6[2] 给定信息表 $K=\langle U,C,V,f\rangle$ 和个体对象集合 X,在非对称相似关系下 X 的上近似和下近似是对在容差关系下 X 的上近似和下近似的改进。

结束语 本文是矩阵论方法在基于粗糙集理论知识获取中的具体应用,其主要目的是运用矩阵这一有力的数学工具,对粗糙集的基本概念和基本运算性质给出一种较为系统和完整的描述。在特征矩阵逻辑运算中,同时定义"与积"和

"或积"两种运算，较好地实现了上述目的，尤其在不同关系性质条件下，粗糙集理论中的上、下近似计算和对象关系类的矩阵表示定理，为基于粗糙集理论的知识表示与知识获取提供了一种能行与可计算的思路与方法。矩阵论方法在粗糙集理论中应用的意义在于它具有一般性，如：通过定义特征向量的"模"，就可以对粗糙集的精度进行表示与计算，此外，矩阵论方法还可推广到与模糊粗糙集和粗糙模糊集有的基本概念与基本运算的表示，并为之提供能行有效的算法。

感谢 高尚博士为本文提供资料和成文过程中的积极建议。

参 考 文 献

1 Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341～356
2 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
3 雷晓蔚. 粗集理论的矩阵方法[J]. 计算机工程与应用，2006，42(17)：73～75
4 李龙星，运士伟，杨炳儒. 粗糙集概念与运算的布尔矩阵表示[J]. 计算机工程，2005，31(14)：16～17
5 仁艳玲，朱明放. 基于粗糙集的属性约简的矩阵方法[J]. 陕西理工学院学报，2006，22(3)：76～80
6 张桂芸，黄国兴，杨炳儒. 基于分辨相似矩阵的相似粗糙集的属性约简算法[J]. 计算机工程，2006，32(10)：43～44
7 高学军，丁军. 基于简化差别矩阵的属性约简算法[J]. 系统工程理论与实践，2006，20(6)：101～107

(上接第 161 页)

宣扬、传播对国家安全有害内容的文本为 1800 篇，它们构成属于类型 c_1 的文本集；揭露、批判这种对国家安全有害内容的文本为 3716 篇，它们构成属于类型 c_2^1 的文本集；内容与那些对国家安全有害的内容完全不同，但它们使用的词语中有相当部分是相同的文本为 828 篇，它们构成属于类型 c_2^2 的文本集；其他文本为 6256 篇，它们构成属于类型 c_2^3 的文本集；文本集 c_2^1、c_2^2 和 c_2^3 共同构成属于类型 c_2 的文本集，共 10800 篇。为了模拟现实环境中两类文本出现的实际情况，属于类型 c_1 和属于类型 c_2 的文本数目比例为 1∶6。将属于类型 c_1 和属于类型 c_2 的文本集随机地平均分为四份，以其中的一份构成测试集，另外的三份构成训练集，按四栏进行交叉验证，以四栏实验的平均值作为最终的性能指标。对文本分类的效率用分类所耗时间来进行评估。实验所用 PC 配置如下：CPU(Intel Pentium4 3.0)、内存(DDRII533 1G)。

3.2 特征的选择

在实验中，以词语作为中文语料的特征，文中选用了清华大学开发的 CsegTag3.0 对中文进行分词。文中采用了改进的互信息公式(12)[3]进行特征选择。

$$MI_1(t_k,c_i)=\sum_{i=1}^{n}P\{t_k,c_i\}\log\frac{P\{t_k,c_j\}}{P\{t_k\}P\{c_i\}} \qquad (12)$$

3.3 实验结果

由于测试集中，属于类型 c_1 和属于类型 c_2 的文档比例为 1∶6，如果将所有文本都标记为 c_2，类型 c_2 的分类精度也能达到 85.7%，因此类型 c_2 的分类性能对所选择的分类方法不敏感。为了节约篇幅，在下面的实验中只给出类型 c_1 的分类性能。

实验 1 两个分界常数 *Dist*1 和 *Dist*2 的确定。

错误率和区域百分比两个评估指标，定义如下：

错误率(Error Rate)：

$$ER=\frac{\text{在某区域内错误分类的文本总数}}{\text{数据集中错误分类的文本总数}}\times 100\%$$

区域百分比(Region Per)：

$$RP=\frac{\text{在某区域内的文本总数}}{\text{数据集中的文本总数}}\times 100\%$$

以词语为特征，以改进互信息公式(12)进行特征选择，采用朴素贝叶斯分类器进行分类，确定文本不可靠区间，如表 1 所示。

表 1 给定区域错误率和区域百分比

区域	$14>dist>0$	$0\geqslant dist\geqslant -28$	$14\geqslant dist\geqslant -28$
错误率	28.56%	66.85%	95.21%
百分比	2.08%	34.79%	36.87%

从表 1 可以看出，不可靠区域为 $-28\leqslant dist\leqslant 14$，95.21% 的错误出现在该区域内，而该区域文本总数仅占所有文本的 36.87%。

实验 2 三种分类模型的性能比较

确定阈值 α 和 β 时，遗传算法初始参数设定参考文[4]中给出的经验值，这些经验在一定程度上具有一定的代表性。上一个实验已经求得二维文本空间中不可靠区域为：$-28\leqslant dist\leqslant 14$，*Dist*1 和 *Dist*2 分别为 14 和 -28，在该不可靠区域内利用遗传算法获取最优 α 和 β，三种分类方法性能比较如表 3 所示。

表 3 在中文语料中，三种分类方法性能的比较

Classifier	SVM	Bayesian	OP-Bayesian
α		1	1.16
β		1	0.93
Precision	99.37%	93.35%	97.98%
Recall	88.94%	88.78%	91.05%
F1	93.85%	91.00%	94.39%
Term Num	1000	500	500
time	23154ms	5241ms	6823ms

结论 本文提出了一种基于朴素贝叶斯和遗传算法的两类文本分类方法，该方法利用文本特征估算文本属于两种类型的测度 X 和 Y，以 X 为横坐标、Y 为纵坐标构造二维文本空间，将文本映射为二维空间中的点，将分类器变换为二维空间中的分割直线；利用遗传算法在此二维空间不可靠区域中寻求一条符合语料集分布的最优分割直线，从而使分类器达到最佳性能。在由 12600 篇文本构成的中文语料数据集上的实验表明，该方法具有较高的分类性能和效率。

参 考 文 献

1 Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002, 34(1): 1～47
2 樊兴华，孙茂松. 一种高性能的两类中文文本分类方法. 计算机学报，2006，29(1)：124～131
3 Sahami M, Dumais S, Hecherman D, Horvitz E. A Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization: Papers from the AAAI Workshop, Madison Wisconsin: [AAAI Technical Report WS-98-05]. 1998. 55～62
4 王小平，曹立明. 遗传算法-理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002. 189～200