文章编号: 1007-4619(2005)04-0398-07

基于粗集的环境机制发现模型及其渔业应用

苏奋振[1],周成虎[1],史文中[2],杜云艳[1],樊 伟[3]
(1.中国科学院 地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101;
2.香港理工大学 土地测量系,香港九龙; 3.农业部海洋与河口渔业重点开放实验室,上海 200090)

摘 要: 地学事件或地学变量受控于环境因子,其关系常为非线性。另一方面,影响变量取值或事件发生的时空范围及其环境要素具有不确定性。环境因子的时空配置关系集中体现这种关系的复杂度。这使得寻找决定事件发生或某些地学变量取值的环境因子及其组合存在困难。针对渔场形成的环境机制发现,构建 RS-STAMM 模型,将时空离散化,以邻域方法提取空间环境变量,形成决策表,利用粗集约简方法,对环境因子及其时空配置关系进行筛选,进而寻找影响事件或变量取值的环境因子的时空配置结构。最后以发现渔场形成的环境机制为目标,将模型应用在渔业遥感研究中,以海洋鱼类聚集的温度场配置提取为实例,验证模型有效性。

关键词: 海洋地理信息系统(MGIS);渔业遥感;关联规则; 渔场;知识发现
中图分类号: P208 **文献标识码**: A

1 引 言

地学世界的复杂性和人们知识的不完全或不精确性,往往导致确定的数理概念或模型在处理不确定事物或事件时有一定的局限性。由此引入不精确概念、离散模型显得尤为必要。

利用一个或多个要素场来估计某事件的发生或另一要素场,常采用回归分析以建立变量间的函数关系。然而回归分析的前提是因变量之间独立和正态分布,这对于环境变量不易达到[1]。回归分析避免变量间的相关性,而地学问题的空间相关往往是我们所关心的。有时变量间的关系也并非可以用确定方程来表示,这在变量类型是类别或序号时更是如此。

更重要的是,地学状态或事件除与其当前位置的环境变量有关外,还与其位置周边环境变量的配置有关[2]。处理哪些位置的变量会影响事件的发生或周边变量如何配置时会影响事件的产生、规模等问题时,往往存在选取变量或变量参数化的困难[3]。

基于此,本文构建了基于粗集的环境因子时空配置提取模型(Rough-Set-Based Spatiotemporal Assignment Mining Model: RS-STAMM)。模型采用离散化思想将空间要素场栅格化,将栅格单元要素场及其相互关系作为先决条件,目标事件或状态作为决策条件。然后利用粗集方法筛选出对目标事件或状态起作用的环境因子及其组合,并确定各自对目标的影响。进而提取决定地学事件或状态的环境因子时空配置。

2 基于粗集的环境因子时空配置提取模型(RS-STAMM)

2.1 邻域与决策表

为便于表述,将状态的取值、地物或现象的产生等,暂称为事件。设定某空间范围的事件为研究对象,称此地域为研究焦点。利用定性的、模糊的背景知识,确定可能影响处于研究焦点事件的空间范围,此范围为邻近范围。焦点和邻近范围构成邻域,如图 1,其中灰色区域为焦点。

邻域的形状、范围及焦点的位置可根据研究需要而定。邻域是获取变量的一个空间模板,从邻域

收稿日期: 2004-01-13;**修订日期**: 2004-03-17
基金项目: 863项目:"中国海岸带及近海卫星遥感综合应用系统技术"(2003AA604040);"大洋金枪鱼渔场渔情速预报技术"(2003AA637030);农业部海洋与河口渔业重点开放实验室开放基金(开-2-04-12)。

作者简介: 苏奋振(1972—),男,副研究员,1994年毕业于武汉测绘科技大学摄影测量与遥感专业,2001获中国科学院地理研究所博士学位。现从事资源环境生态的遥感与 GIS研究,发表论文 30余篇。E-mail: sufz@ lreis.ac.cn。

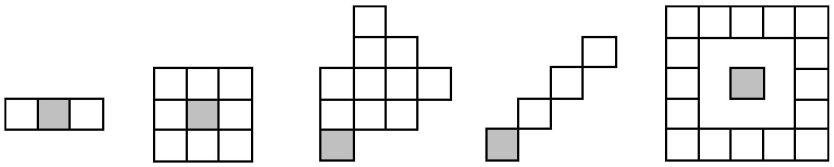

图 1　各种形状的邻域

Fig 1　The structure of neighbour

中获取的环境变量值,可以取邻域中每个单元的平均值,也可以取交叉点的值。获取的环境变量可以是多要素,也可以是单一要素。

那么对应于每一时刻,对应于每一焦点事件的属性,有一组环境变量值与之对应。设影响事件属性的要素集合为 $H=\{h_1, h_2 \cdots h_{k-1}, h_k\}$,则可构建决策表。其中从邻域取了 m个位置上的 h_i值,n个位置上 h_j值,取邻域单元间 h的差值若干,用于描述空间上某种能量差。这些所取变量即为可能影响焦点事件属性 D的条件属性 C。由于 C是根据定性的、模糊的背景知识获得,或者根本就不具备知识,取了一个比较大的邻域范围,其中有些位置上的所有环境要素的值或部分环境要素的值并不会对焦点的事件造成影响。换句话说,焦点事件仅受其中某些位置某些要素值的影响,可能是单独影响,也可能是在一定的配置关系下才产生影响。

由此,需要对决策表中的时空条件属性进行约简,从而获取真正影响焦点事件的环境因子及其时空上的配置关系。为了方便起见我们定义每条记录为一个时空事件。

2.2　基于粗集的约简方法

Rough集理论[4]是一种新的处理模糊性和不精确性知识的数学工具,Rough集理论已经在信息系统分析、人工智能、决策支持系统、知识与数据发现、模式识别与分类、故障检测等方面取得了较为成功的应用[5-7]。我们的模型中采用它来进行约简。

2.2.1　客观世界的表达与不分明关系

表 1是对客观世界的一种表达,记为 $(U, C\cup D, V, f)$,其中 U是一个有限的非空时空事件集,称为全域,C是条件属性集合;D是判别属性集合;V是属性 $C\cup D$的值域集合;函数 $f: U\times C\cup D\rightarrow V$定义对象的属性值。设有 x, y两时空事件,若 f(x)=f(y),则 x与 y为等价类。其中属性 $C\cup D$构成对象的属性空间 A,那么有如下性质:

对于属性空间 A的每一个子集 B都有一个与之相关的二元关系,称为不分明关系,定义如下:

$$IND(B)=\{(x, y)\in U: a\in B, a(x)=a(y)\}$$

其中 a(x)表示对象 x在属性 a上的值。也就是说光靠属性集 B无法区别对象 x与 y,或者说 x, y的差别不是因为属性集 B的不同引起的。

举例说明,设有客观世界 $k=\{U, \{p, q, r\}\}$其中 $U=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$八时空事件组成,U/p表示全域根据 p所得的分类,即 IND(p)。有 $U/p=\{\{x_1, x_5\}, \{x_2, x_4, x_8\}, \{x_3, x_7\}, \{x_6\}\}$, $U/q=\{\{x_1, x_2, x_5, x_7\}, \{x_3\}, \{x_4, x_6, x_8\}\}$, $U/r=\{\{x_1, x_5, x_7\}, \{x_2, x_6\}, \{x_3\}, \{x_4, x_8\}\}$,则用 p分得的第一类 $[X_1]_p=\{x_1, x_5\}$,用 q分得的第一类 $[X_1]_q=\{x_1, x_2, x_5, x_7\}$,若 $R=\{p, q\}$,则 $[X_1]_R=\{x_1, x_5\}$,即 x_1, x_5对于关系 p, q来说都是不可分的,且 $IND(R)=\{\{x_1, x_5\}, \{x_2\}, \{x_3\}, \{x_4, x_8\}, \{x_6\}, \{x_7\}\}$。

2.2.2　近似集

存在类别 $X\subseteq U$,且 $X\neq\phi$。前面例子中 U/R为若干个子集 X_i, $i=1, 2, \cdots, n$, X_i称为 R一基本集。则可用 R基本集来描述 X。可以将那些包含在 X中的 R的基本集的并,定义为 X的 R正区域,记为 $R\_(X)=\{x\in U \mid [X_i]_R\subseteq X\}$;将那些所有与 X的交非空的 R基本集的并,定义为 X的 R负区域。

沿用上面的例子,$R=\{p, q, r\}$则 $IND(R)=\{\{x_1, x_5\}, \{x_2\}, \{x_3\}, \{x_4, x_8\}, \{x_6\}, \{x_7\}\}$。对于 U上的子集 $X=\{x_2, x_4, x_7\}$可得到 $R\_(X)=\{x_2\}\cup\{x_7\}=\{x_2, x_7\}$, $R^-(X)=\{x_4, x_8\}\cup\{x_2\}\cup\{x_7\}=\{x_4, x_2, x_8, x_7\}$。

2.2.3　约简与依赖

为了从众多的时空条件属性中提出必要的时空配置,引入约简与依赖的概念。

设属性 $r\in R$,若 $IND(R)=IND(R-\{r\})$,则称属性 r在 R中是可省的(dispensable)。否则就是不可省的。若属性集 R中的每个属性或属性组合都是不可省的,则称 R是独立的(independent),否则是依赖的或非独立的,也就是可省的。

进一步可定义,若存在属性集 Q和 P, $Q\subseteq P$, Q

是独立的,若存在 IND(Q)=IND(P),则称 Q是属性集 P的一个约简(reduct),在属性集 P中所有不可省的属性的集合称为 P的核(core),以 core(P)来表示,可见,对于一个属性集 P,一般均有多个约简,且 core(P)=∩ reduct(P)。以上面的例子为例,若 R={p, q, r}则

IND(R) = {{x_1, x_5}, {x_2}, {x_3}, {x_4, x_8}, {x_6}, {x_7}}

IND(R-{r}) = {{x_1, x_5}, {x_2}, {x_3}, {x_4, x_8}, {x_6}, {x_7}} = IND(R)

IND(R-{p}) = {{x_1, x_5, x_7}, {x_2}, {x_3}, {x_4, x_8}, {x_6}} ≠ IND(R)

IND(R-{q}) = IND(R)

则 p不可省,q, r可省。如此,由 {p, q, r}三个等价关系组成的集合 {p, q, r}与 {p, q}, {p, r}定义了相同的不分明关系(类),又 IND({p, q}) ≠ IND({p}), IND({p, q}) ≠ IND({q}),则 {p, q}和 {p, r}就是 R的简化,而且 {p}是 R的核。

对于上表,若存在条件属性集 B⊂C,则可以定义 D的 B正区域 $POS_B(D)$为

$$POS_B(D)=U\{B_-(X)\mid X\in IND(D)\}$$

即 D的 B正区域是全域 U上的所有那些使用 B进行划分的类能够正确地分类于用 D划分的类的集合。举例说明,设 IND(D)=U/D={{x_1, x_2, x_3, x_4, x_8}, {x_5, x_6, x_7}}, U/B={ {x_1, x_5}, {x_2}, {x_3}, {x_4, x_8}, {x_6}, {x_7}},则 $POS_B(D)$={x_2} U {x_3}· U {x_4, x_8} U {x_6} U {x_7}={x_2, x_3, x_4, x_8, x_6, x_7}。

如此可以引出属性集的依赖性:

$$k=\gamma_B(D)=card(POS_B(D))/card(U)$$

即属性集 D以依赖度 k($0\leqslant k\leqslant 1$)依赖于属性集 B,或者说属性集 D的取值在多大程度上决定于 B的取值。其中 card描述集合中的元素个数。对应于决策表,card为记录个数。若 k=1,则称属性集 D完全依赖于 B;若 $0<k<1$,则称属性集 D部分依赖于属性集 B;若 k=0,则称属性集 D完全独立于属性集 B。

2.3 模型流程

为了便于说明模型的运行,我们给出模型的流程图(图 2)。首先是将研究的时空过程在时间上分解成一序列时刻的空间状态。对每一个空间状态,利用先验知识,将可能影响研究对象的空间范围、尺度、拓扑关系、距离关系、方向关系等转换为邻域概念。利用邻域从空间状态中获取参数,参数的选取也可以参考先验知识。如此形成决策表。对连续的属性值进行区间化,以减少运算次数。对决策表进行约简,获取主则。对规则进行分析,获取知识,修改邻域的定义或合并若干连续值的区间。不断的反馈,直到获取满意的规则为止。

从流程图 2可以看出,模型能够对初始的先验知识进行不断的修正。在先验知识不够的情况下,可以将邻域的范围和关系定义多些,邻域单元或尺度小些,环境要素多些;在连续值区间化时,区间小些。经过一次运算,根据所获得的规则对邻域的大小、尺度、关系进行调整,舍弃在规则集中不出现的要素,合并属性区间相临的规则,进而在第二次迭代时合并这些连续值区间。

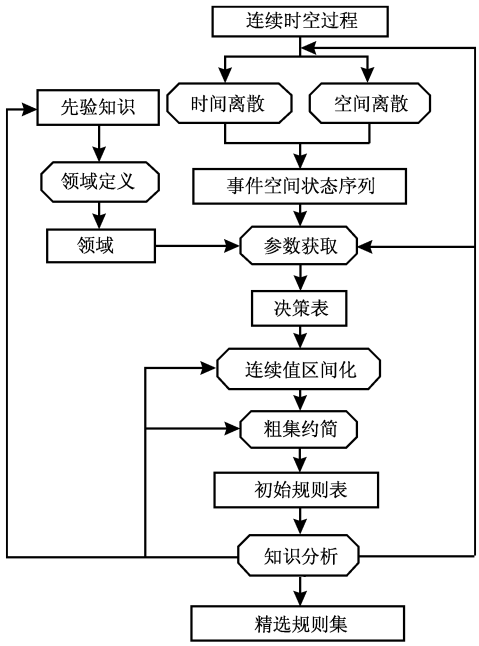

图 2 模型流程图

Fig 2 Flow chart of model

由此可见模型的运算量和迭代次数与先验知识是一对反比关系,即先验知识越精确,运算量就愈小。先验知识越模糊,计算量就越大。

3 栅格化空间结构提取实例

由于海洋鱼类对环境有一定趋向性,海洋要素场空间上的配置不同,制约着海洋鱼类的分布,从而"调制"区域的鱼群密度。而哪些位置的要素和它们的何种空间配置或时空配置决定鱼群的集聚,就是本例要探究的。如何从多年数据中获取渔场形成的环境时空配置关系,传统方法存在一定的限制。

以下用本文方法探求海洋要素场的时空配置如何“调制”鱼群密度。为方便叙述,仅以遥感反演的温度场为例。

渔业传统上将海区按 30′×30′的经纬格网进行划分,每一格网为一个渔区。如图 3所示,选取 $T_6T_7T_aT_b$四角点构成的渔区为研究对象。为了考察温度场对该渔区内鱼类聚集的影响,我们从温度场上取一定邻域上点的温度,见图 3的小方点,为了方便展示计算过程和易读性,只列部分数据如表 1。时间上也只取每年的第一周,表中周次 8701表示 1987年第一周。事实上,可以取尽量多点上的值及其相互关系,通过约简算法,无关的参量将被自动摒弃。时间上,也可在一次搜索中,遍历所有时间。

表中 T_g-T_1表示邻域中东南 西北方向的梯度,WC表示该渔区该周的上层鱼平均网产。根据先验知识,由于该区域渔场受黄海暖流进退影响,而其行进方向为东南 西北方向,由于每增加一个条件属性,计算量指数增加,考虑篇幅,在此温度梯度仅取该方向梯度。

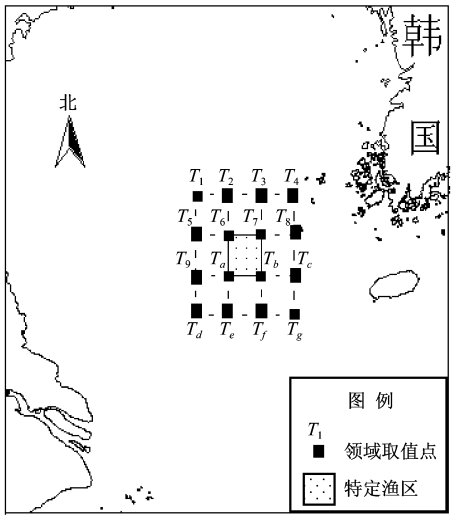

图 3　特定渔区及其邻域定义

Fig 3　The fishing cell with its neighbour

表 1　特定渔区渔场形成事件表格在数据库中的组织

Table 1　Fishing Ground in the Fishing Cell

周次	T_1	T_6	T_7	T_b	T_a	T_g	T_g-T_1	WC
8801	11.5	14.2	14.1	17	15.1	18	6.5	0
8901	11.7	11.2	12.1	12.3	11.3	13.8	2.1	4170
9001	13.3	14	12	15	14.7	16	2.7	1275
9101	13	13.1	13	13.8	14.8	15	2.0	2650
9201	14.2	13.7	13.7	14.3	14.6	15.2	1	0
9301	13	13.9	12.9	14	14	17	4	1112
9401	12.3	14.8	15	14	13.4	17.1	4.8	740
9501	12.5	12.7	12.3	13	14	16.4	3.9	0
9601	12.7	12.6	11.6	13.5	13.1	14.6	1.9	0
9701	11.5	13	13.9	13.9	11.7	16.5	5	0

从表 1难以看出温度与渔场间的关系,即使把温度和网产展布在图上,也看不出规律(限于篇幅略去图),更何况从数据仓库中发现知识是针对海量数据的,受时间和记忆力的限制,人工比较无法胜任。同样如表 2,经 Spearman相关检验,T_1, T_6, T_7, T_b, T_a, T_g, T_g-T_1与 WC相关置信度都大于 0.05,即不相关。事实上温度是影响该区域渔场形成的主要因素。这说明一般的相关分析或曲线拟合不适于这种应用。究其原因是渔场的形成只存在于一定的温度范围内和一定的温度梯度范围内,而在这范围内不存在共同趋势关系。

为此,采用本文的方法处理。首先对表 1的连续型数值作区间化,将数值型转变为布尔型。区间化与前面邻域选取都是非常关键的步骤。若有一定的背景或定性认识,将有助于合适的选取。反之,可利用迭代反复尝试,或将提取的细小相临规则的区

表 2　特定渔区邻域温度值与渔场密度(WC)Spearman秩相关系数

Table 2　Spearman's Correlation coefficients between Point Temperature and the Fish Density

	T_1	T_6	T_7	T_b	T_a	T_g	T_g-T_1	WC
与 WC的相关系数	0.202	−0.032	0.265	−0.238	−0.032	0.369	−0.187	1.000
置信度	0.576	−0.929	0.459	0.565	0.929	0.295	0.604	

间进行合并，直到所提取的规则没有相邻区间，或支持度足够大，规则合理为止[2]，即模型流程的反馈机制。

对温度值离散化：A: <12.9; B: 13.0—15.0; C: >15.1

对平均网产离散化：有：WC>500箱/网；无：WC$\leqslant 500$箱/网（箱$=20$kg）

T_g-T_1值离散化：Y: $T_g-T_1\geqslant 2$; N: $T_g-T_1<2$，则表 1转换为表 3。

表 3 特定渔区渔场形成各变量区间化

Table 3 Discrete Value in the Fishing Cell

周次	t_1	T_6	T_7	T_b	t_a	t_g	T_g-T_1	WC
8801	A	B	B	c	c	c	Y	无
8901	A	A	A	a	a	b	Y	有
9001	B	B	A	b	b	c	Y	有
9101	B	B	B	b	b	b	Y	有
9201	B	B	B	b	b	c	N	无
9301	B	B	A	b	b	c	Y	有
9401	A	B	B	b	b	c	Y	有
9501	A	A	A	b	b	c	N	无
9601	A	A	A	b	b	b	Y	无
9701	A	B	B	b	a	c	Y	无

表 1中每条记录记为一个时空事件，表达在什么地方什么时间发生了什么的信息。则表中包含10个事件，即 e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}。其中 T_1, T_6, T_7, T_b, T_a, T_g, T_g-T_1 为其空间相互关系和非空间属性，记 $C=\{T_1,\cdots,T_7\}$为条件属性，WC为决策属性 $D=\{WC\}$，则有

$IND(C)=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

$IND(C-\{T_1\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

$IND(C-\{T_6\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

$IND(C-\{T_7\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

$IND(C-\{T_b\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

$IND(C-\{T_a\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7,e_{10}\},\{e_8\},\{e_9\}\}$

$IND(C-\{T_g\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

$IND(C-\{T_g-T_1\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4\},\{e_5\},\{e_7\},\{e_8\},\{e_9\},\{e_{10}\}\}$

因为 $IND(C)\neq IND(C-\{T_a\})$可见 T_a 为不可省属性，其它为可省属性。

$IND(\{WC\})=\{\{e_1,e_5,e_8,e_9,e_{10}\},\{e_2,e_3,e_4,e_6,e_7\}\}$

因为 T_a 为不可省，故二项集中必有 T_a，接下来看两个属性的组合。

$IND(\{T_a,T_1\})=\{\{e_1\},\{e_2,e_{10}\},\{e_3,e_4,e_5,e_6\},\{e_7,e_8,e_9\}\}$

$POS^{(\{WC\})}_{\{T_a,T_1\}}=\{e_1\}$

$\gamma^{(\{WC\})}_{\{T_a,T_1\}}=1/10$

$IND(\{T_a,T_6\})=\{\{e_1\},\{e_2\},\{e_3,e_4,e_5,e_6,e_7\},\{e_8,e_9\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a,T_6\}}=\{e_1,e_2,e_8,e_9,e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a,T_6\}}=5/10$

$IND(\{T_a,T_7\})=\{\{e_1\},\{e_2\},\{e_3,e_6,e_8,e_9\},\{e_4,e_5,e_7\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a,T_7\}}=\{e_1,e_2,e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a,T_7\}}=3/10$

$IND(\{T_a,T_b\})=\{\{e_1\},\{e_2\},\{e_3,e_4,e_5,e_6,e_7,e_8,e_9\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a,T_b\}}=\{e_1,e_2,e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a,T_b\}}=3/10$

$IND(\{T_a,T_g\})=\{\{e_1\},\{e_2\},\{e_3,e_5,e_6,e_7,e_8\},\{e_4,e_9\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a,T_g\}}=\{e_1,e_2,e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a,T_g\}}=3/10$

$IND(\{T_a,T_g-T_1\})=\{\{e_1\},\{e_2,e_{10}\},\{e_3,e_4,e_6,e_7,e_9\},\{e_5,e_8\}\}$

$POS^{(\{WC\})}_{\{T_a,T_g-T_1\}}=\{e_1,e_5,e_8\}$

$\gamma^{(\{WC\})}_{\{T_a,T_g-T_1\}}=3/10$

因为 $\gamma^{(\{WC\})}_{\{T_a,T_6\}}$最大，即两个属性组合中，决策属性最依赖于 $\{t_a,t_6\}$，接下来看三个属性的组合。

$IND(\{T_a,T_6,T_1\})=\{\{e_1\},\{e_2\},\{e_3,e_4,e_5,e_6\},\{e_7\},\{e_8,e_9\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a,T_6,T_1\}}=\{e_1,e_2,e_7,e_8,e_9,e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a,T_6,T_1\}}=6/10$

$IND(\{T_a,T_6,T_7\})=\{\{e_1\},\{e_2\},\{e_3,e_6\},\{e_4,e_5,e_7\},\{e_8,e_9\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a,T_6,T_7\}}=\{e_1,e_2,e_3,e_6,e_8,e_9,e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a,T_6,T_7\}}=7/10$

$IND(\{T_a,T_6,T_b\})=\{\{e_1\},\{e_2\},\{e_3,e_4,e_5,e_6,e_7\},\{e_8,e_9\},\{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a, T_6, T_b\}} = \{e_1, e_2, e_8, e_9, e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a, T_6, T_b\}} = 5/10$

$IND(\{T_a, T_6, T_g\}) = \{\{e_1\}, \{e_2\}, \{e_3, e_5, e_6, e_7\}, \{e_4\}, \{e_8\}, \{e_9\}, \{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a, T_6, T_g\}} = \{e_1, e_2, e_4, e_8, e_9, e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a, T_6, T_g\}} = 6/10$

$IND(\{T_a, T_6, T_g - T_1\}) = \{\{e_1\}, \{e_2\}, \{e_3, e_4, e_6, e_7\}, \{e_5\}, \{e_8\}, \{e_9\}, \{e_{10}\}\}$

$POS^{(\{WC\})}_{\{T_a, T_6, T_g - T_1\}} = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}\}$

$\gamma^{(\{WC\})}_{\{T_a, T_6, T_g - T_1\}} = 10/10$

从结果来看,WC完全依赖于 $\{T_a, T_6, T_g - T_1\}$,也就是空间配置 $\{T_a, T_6, T_g - T_1\}$决定 {WC},或者说,是否形成渔场取决于左边的温度,与及东南与西北的温差。根据前面对温度的分段,显然该渔区要形成渔场的条件是左边两角的温度处于 13℃－15℃间且东南与西北温差大于 2℃,则形成渔场。

从此可以看出对于此渔区,在每年的第一周,若其周围四点中左上点海温为 13℃－15℃,左下点温度也在 13℃－15℃,同时东南 西北方向存在一定的温差,也就是西北温度低而东南温度高,这样鱼类活动空间被压缩在较小范围内,从而形成高密度聚集。

需要说明的是,考虑到篇幅和运算的明晰性,文中没有用所有时间和所有渔区的数据,事实上,时间和空间位置也是条件属性之一,经过模型的运算,其与其它属性的配置关系将被提取出来。我们对东海水域的温度和渔场数据,进行了时空关联规则的提取,并将规则用于后续的专家系统知识库中,利用遥感反演的海面温度场来预报渔场,其有效率达 80%。

4　讨　论

基于离散化和邻域的思想构建了环境因子及其配置结构的提取模型。模型充分考虑了事件发生中其相邻区域环境因子及其配置关系对其所产生的影响,从而补充了多元相关分析,类比推理等方法的某些不足。模型除可以用于寻找影响动物行为的环境时空配置外,也可以尝试用于发现地理状态分布与环境要素的时空配置的关系,比如寻找地价与环境的关系,居民点与环境的关系,城市大小与环境的关系等。

栅格化有利于直接利用遥感数据,邻域的构建能够将拓扑关系、距离关系、方向关系以及属性值的相对关系考虑进去。

实例中用该模型提取的温度场与渔场的时空关联规则,也可认为是提取了从环境场诊断出渔场的知识,即提取诊断规则(知识)。所发现的知识可用于预报渔场发生的专家系统,又或可作为案例推理的案例,也可以作为指导渔业生产或提供渔业资源环境研究的先验事件。

参 考 文 献 (References)

[1] Wang J F, Li L F, Ge Y, et al. A Theoretic Framework for Spatial Analysis[J]. Acta Geographica Sinica, 2000, **55**(1): 92－103. [王劲峰,李连发,葛咏等. 地理信息空间分析的理论体系探讨[J]. 地理学报, 2000, **55**(1): 92－103.]

[2] Su F Z, Zhou C H, Liu B Y, et al. A Spatiotemporal Pattern Extracting Model for Fishing Ground[J]. Acta Oceanologica Sinica. 2002, **24**(5): 44－56. [苏奋振,周成虎,刘宝银等. 基于海洋要素时空配置的渔场形成机制发现模型和应用[J]. 海洋学报, 2002, **24**(5): 44－56.]

[3] Su F Z, Zhou C H, Vincent Lyne, et al. A Data Mining Approach to Determine the Spatio-temporal Relationship Between Environmental Factors and Fish Distribution[J]. Ecological modeling, 2004, **174**: 421－431.

[4] Pawlak Z. Roughsets Z[M]. Norwell: Kluwer Academic Publishers. 1991.

[5] Zheng Z K, Zhang G F, Shao H H. Data Mining and Knowledge Discovery: an Overview and Prospect[J]. Information and Control, 1999, **28**(5): 21－24. [郑之开,张广凡,邵惠鹤. 数据采掘与知识发现:回顾和展望[J]. 信息与控制, 1999, **28**(5): 21－24.]

[6] Di K C, Li D R, Li D Y, et al. Rough Set Theory and Its Application in Attribute Analysis and Knowledge Discovery in GIS[J]. Journal of Wuhan Technical University of Surveying and Mapping. 1999, **24**(1): 6－10. [邸凯昌,李德仁,李德毅等. Rough集理论及其在 GIS属性分析和知识发现中的应用[J]. 武汉测绘科技大学学报, 1999, **24**(1): 6－10.]

[7] Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases[A]. Advances in Spatial Databases[C]. Proceedings of 4th Symposium, SSD'95. Berlin: Springer-Verlag, 1995, 47－66.

Rough-Set-Based Spatiotemporal Assignment Mining Model with Its Application for Marine Fishery

SU Fen-zhen[1], ZHOU Cheng-hu[1], SHI Wen-zhong[2], DU Yun-yan[1], FAN Wei[3]
(1. LREIS, Institute of Geography Science and Natural Resources Research, CAS, Beijing 100101, China;
2. Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China;
3. East China Sea Fisheries Research Institute, CAFS, Shanghai 200010, China)

Abstract: Geo-event is controlled by the environment factors with nonlinear relationship. That means it is important to discover the spatiotemporal assignment of environmental factors. Vector-based association rule discovery models have been provided to look for environmental pattern successfully in terrestrial applications recently. But they just consider the topological relationship in static state between parcels or objects with complex algorithm. And the data type of the attribution has to be limited as category. Except the topologic relationship, it is difficult to take other relationships into account. In the environment research, the continue field also is important research content, especially in meteorology and oceanography. And the continue field always includes the temporal property as geographic phenomenon or geographic progress. This work is to deal with the process. It disperses the space with time. A neighborhood is defined to extract the spatiotemporal variables to make a decision table. A reduction algorithm based on rough set will mine the spatiotemporal assignment with the environmental factors. Finally, the model was applied in fishery geography to find the environmental pattern, which determines the form of fishing ground.

Key words: marine geographical information system (MGIS); fishery remote sensing; spatiotemporal association rule; fishing ground; knowledge discovery