

空间遥感数据的多粒度标记分类

吴伟志 杨玉芳

(浙江海洋学院数理与信息学院 舟山 316004)

摘要 以空间遥感数据为研究对象,研究多粒度标记数据的分类问题。首先,根据地表物的不同物理含义,可以得到一个多粒度的决策表。然后,通过知识约简发现隐藏在多粒度决策表中的 if-then 分类规则。最后,根据实际的遥感数据给出分类的精度。

关键词 分类,遥感,多粒度,粒计算,粗糙集

中图法分类号 TP18 **文献标识码** A

On Multi-granular Labeled Classification for Spatial Remote Sensing Data

WU Wei-zhi YANG Yu-fang

(School of Mathematics, Physics and Information Science, Zhejiang Ocean University, Zhoushan 316004, China)

Abstract This paper studied the discovery of classification rules for remote sensing data with multi-granular labels. According to the different physical meaning of land cover, we first obtained a multi-scale decision table, and then discovered the classification rules hided in the multi-scale decision table. Finally, we analyzed the accuracy of the classification.

Keywords Classification, Remote sensing, Multi-scale, Granular computing, Rough set

1 引言

粗糙集理论^[1,2]能够分析和处理不精确、不协调和不完备信息,是由波兰数学家 Pawlak 于 20 世纪 80 年代初提出的用于数据分析的理论,该理论作为一种具有极大潜力和有效的知识获取工具受到了人工智能研究工作者的广泛关注。基于粗糙集理论的知识获取的核心问题之一是属性约简,人们从知识库中获取的知识只有在约简的基础上才有推广和应用的价值。由于粗糙集方法在属性约简和规则提取方面的独特优势,该方法已经被成功地应用于地理信息系统和空间遥感数据分析^[3-8]。

在实际问题中,从数据库中提取学习规则集往往依赖于数据的背景知识。在遥感图像分类问题中,地表物(如植被)的数据是以不同的光谱波段反射的灰度值来显示的。这类数据的分类任务是选择一个小的光谱波段集合使得与原来的光谱波段集合具有相同的分类能力。需要指出的是,地表物的光谱波段的反射灰度值是取区间 $[0, 255]$ 中的一个实数。因此,原始数据(训练集)是一个实值信息系统,但是将这样的实值信息系统作为训练样本,所得到的规则集过于庞大,其推广能力很差。因此,要将实值信息系统进行适当变换或做离散化处理。与模糊集方法和神经网络方法等其它传统方法^[9-18]不同,Leung 等^[19,20]提出了粗糙集与统计相结合的数据分析处理模型。该方法先用统计方法将实值信息系统变换成为一个区间值信息系统用于处理遥感图像分类问题,其基于如下假设:同一个地表物在每一个光谱波段上的反射灰度值服从

正态分布,从而可以将每一类地表物的实值信息转化为 $\mu \pm 2\sigma$ (μ 是均值, σ 为方差) 的区间值,于是将原始的样本数据集转化为一个区间值的信息系统。然后提出了一个粗糙集模型给出针对空间地表物的遥感数据分类问题。

众所周知,人们在处理复杂问题进行数据分析时,一般会根据实际背景,凭借已有的经验或专业背景知识,将问题分解,然后从不同层次和角度进行分析和计算。从不同的粒度上观察、分析与解决同一问题,是人类智能的特点之一。近年来,作为知识表示和问题求解的工具之一,粒计算(Granular Computing, GrC)的研究获得了快速发展。粒计算是模拟人类思考问题的自然模式,以粒为基本计算单位,处理大规模复杂数据集和信息等建立有效的计算模型为目标。尽管已经提出了很多粒计算模式^[21-23],但是对于同一对象的信息多粒度表示和多粒度问题求解尚缺乏具体的模型。最近, Wu 和 Leung 在文献^[24]中提出了多尺度决策表中基于粒度标记的粗糙集数据分析理论。

本文用粗糙集理论提出空间遥感数据的多粒度标记分类方法,并给出相应粒度下的分类精度。

2 区间值信息系统中的知识发现

2.1 Pawlak 信息系统的概念

一个完备 Pawlak 信息系统 S 是一个二元组 (U, A) , 其中 $U = \{x_1, x_2, \dots, x_n\}$ 是非空有限对象集,称为论域, $A = \{a_1, a_2, \dots, a_m\}$ 是非空有限属性集,使得 $a: U \rightarrow V_a, a \in A$, 即对于 $x \in U, a(x) \in V_a$, 其中 V_a 为属性 a 的值域。

到稿日期:2011-05-20 返修日期:2011-09-02 本文受国家自然科学基金项目(61075120, 61173181)资助。

吴伟志(1964—),男,博士,教授,博士生导师,主要研究方向为粗糙集、粒计算、数据挖掘、近似推理, E-mail: wuwz@zjou.edu.cn; 杨玉芳(1986—),女,硕士生,主要研究方向为遥感数据分析、粗糙集。

一个决策表是一个特殊的信息系统 $S=(U, A \cup \{d\})$, 其中 $d, d \notin A$, 称为决策属性, A 称为条件属性集。可以将一个决策属性看成一个在对象集上的分类器。在机器学习研究中, 决策表也称为训练样本集。

不失一般性, 假设 $V_d = \{1, 2, \dots, I\}$, 则决策属性 d 可以确定论域上的一个划分:

$$U/d = \{[x]_d : x \in U\} = \{X_1^d, X_2^d, \dots, X_I^d\}$$

式中, $X_i^d = \{x \in U : d(x) = i\}, i = 1, 2, \dots, I$ 。集合 X_i^d 称为决策表 $S=(U, A \cup \{d\})$ 的第 i 个决策类。这样, i 可以看成是 X_i^d 的类标记。

2.2 区间值信息系统

一个区间值信息系统是一个二元组 $K=(U, A)$, 其中 $U = \{u_1, u_2, \dots, u_I\}$ 是对象集, $A = \{a_1, a_2, \dots, a_m\}$ 是属性集, 且使得 $a_k(u_i) = [l_i^k, u_i^k], l_i^k < u_i^k, i = 1, 2, \dots, I, k = 1, 2, \dots, m$, 即对象在每一个属性上的取值都是实数区间。

表 1 是一个区间值信息系统, 它是由 5 个地表物和 4 个光谱波段构成的实值信息系统经过统计转换后的遥感数据, 这里 u_i 实际上是地表物的类别。

对于一个区间值信息系统 $K=(U, A), B \subseteq A$, 定义一个 U 上的二元关系, 用 R_B 表示:

$$R_B = \{(u_i, u_j) \in U \times U : a_k(u_i) \cap a_k(u_j) \neq \emptyset, \forall a_k \in B\}$$

两个类 u_i 和 u_j 有关系 R_B , 当且仅当它们不能被属性集 B 区别开来。很明显, R_B 是自反和对称的, 但是不传递。

因此 R_B 是一个相容关系, 满足:

$$R_B = \bigcap_{b \in B} R_{(b)}$$

记

$$S_B(u_i) = \{u_j \in U : (u_i, u_j) \in R_B\}, u_i \in U$$

$S_B(u_i)$ 是 R_B 上 u_i 的相容类, $u_j \in S_B(u_i)$ 当且仅当 u_i 和 u_j 不能被属性集 B 区分开。

表 1 区间值信息系统

U	a_1	a_2	a_3	a_4
u_1	[60, 77]	[44, 68]	[19, 30]	[5, 16]
u_2	[77, 81]	[74, 80]	[53, 60]	[5, 19]
u_3	[74, 97]	[65, 116]	[63, 107]	[56, 107]
u_4	[51, 242]	[73, 275]	[108, 205]	[104, 204]
u_5	[54, 57]	[38, 42]	[94, 139]	[29, 41]

表 2 区间值信息系统

U	a_1	a_2	a_3	a_4
u_{12}	[62, 86]	[44, 89]	[8, 73]	[5, 18]
u_3	[74, 97]	[65, 116]	[63, 107]	[56, 107]
u_4	[51, 242]	[73, 275]	[108, 205]	[104, 204]
u_5	[54, 57]	[38, 42]	[94, 139]	[29, 41]

注意: 这是从实际生活应用中得到的实验结果。简单地说, a_k 是属性, $k = 1, 2, 3, 4, u_j$ 是类, $j = 1, 2, \dots, 5$ 。

表 1 和表 2 是从不同粒度标记下对同一地表区域进行分类的区间值信息系统。相对表 1, 表 2 是粒度粗的信息系统, 它将表 1 中类 u_1 和 u_2 合并标记为同一类, 故表 2 的类比表 1 少, 而且类 u_{12} 的区间对原类 u_1 和 u_2 的覆盖度极高。 a_1 属性的覆盖度 $c_1 = 1 - \frac{62-60}{(81-77)+(77-60)} = 82\%$, 同理可以得到

$$c_2 = 100\%, c_3 = 100\%, c_4 = 1 - \frac{19-18}{19-5} = 93\%。$$

粗糙集理论应用研究的一个重要方向是寻找一个最小的属性子集, 使得该属性子集与属性全集具有相同或相似的分

类能力, 这样的子集叫做属性约简^[1,2]。为了从信息系统得到简洁的决策规则, 需要进行知识约简。这一部分主要研究区间值信息系统的属性约简问题。从而, 可以建立区间值信息系统的分类规则。

假设 $K=(U, A)$ 是一个区间值信息系统, 并且 $B \subseteq A$, 如果 $R_B = R_A$, 那么称 B 是 K 的一个分类协调集。如果 B 是 K 的分类协调集, 而对于所有的 $b \in B, B - \{b\}$ 都不是 K 的分类协调集, 也就是说 $R_{B-\{b\}} \neq R_A$, 那么 B 就叫作 K 的(全局)分类约简。 K 中所有分类约简的集合用 $re(K)$ 表示。所有分类约简的交集叫作 K 的分类核, 分类核中的元素是绝对必要属性, 删除其中的任何一个都将影响分类质量。

另一方面, 对于 $u \in U, B \subseteq A$, 如果 $S_B(u) = S_A(u)$, 那么就称 B 是 K 中类 u 的一个协调属性集。如果 B 是 K 中类 u 的协调属性集, 且对于所有 $b \in B, B - \{b\}$ 都不是 u 的协调属性集, 也就是说 $S_{B-\{b\}}(u) \neq S_A(u)$, 那么 B 叫作 K 中类 u 的(局部)分类约简。 K 中 u 的所有分类约简的集合用 $re(u)$ 来表示。 u 的所有分类约简的交集叫作 K 中 u 的分类核。

需要注意的是, 局部约简不一定必须包含在某一个全局约简内。然而, 如果 B 是一个全局约简, 那么对于任何一个类 u 来说, 一定存在一个 u 的局部约简是 B 的子集。从局部约简能够简化由全局约简推出的分类规则, 可能有更好的推广能力。

一个 K 中的分类协调集是保持所有类的相容类不变的属性子集, 而一个分类约简是保持所有类的相容类不变的最小协调集, 从而会得出相同的分类。剩下的属性是多余的, 删除它们不会对分类有影响, 也就是说不会使分类结果变坏。

下面通过定义可辨识属性集, 用布尔推理方法来计算属性系统的约简。

假设 $K=(U, A)$ 是一个区间值信息系统。记

$$D_{ij} = \{a_k \in A : a_k(u_i) \cap a_k(u_j) = \emptyset\}, i \neq j$$

且 $D_i = \emptyset, i = 1, 2, \dots, I$ 。

D_{ij} 叫作 K 中类 u_i 和 u_j 的可辨识属性集。记

$$M = (D_{ij} : i, j = 1, 2, \dots, I)$$

M 叫作 K 的辨识矩阵。记

$$M_0 = \{D_{ij} : D_{ij} \neq \emptyset\}$$

根据表 1 和表 2, 可以分别得到它们的辨识矩阵, 见表 3 和表 4。因为 $D_{ij} = D_{ji}$, 为简便起见, 只列出 $D_{ij}, 1 \leq j < i \leq I$ 。

表 3 由表 1 得到的辨识矩阵

	u_1	u_2	u_3	u_4	u_5
u_1					
u_2	$a_2 a_3$				
u_3	$a_3 a_4$	$a_3 a_4$			
u_4	$a_2 a_3 a_4$	$a_3 a_4$	a_3		
u_5	A	A	$a_1 a_2 a_4$	$a_2 a_4$	

表 4 由表 2 得到的辨识矩阵

	u_{12}	u_3	u_4	u_5
u_{12}				
u_3	a_4			
u_4	$a_3 a_4$	a_3		
u_5	A	$a_1 a_2 a_4$	$a_2 a_4$	

约简计算可以转化为布尔函数的极小析取范式来计算。已证找一个 Pawlak 信息系统的约简问题可以通过转化为布尔推理的相关问题来解决。在这里把这个方法推广运用到所

提出的区间值信息系统中。

假设 $K=(U,A)$ 是一个区间值信息系统, 信息系统 K 的一个辨识函数(又称辨识公式) f_K 是一个由 m 个布尔变量 $a_1, a_2, \dots, a_m$ 构成的布尔函数, 分别对应属性 $a_1, a_2, \dots, a_m$, 定义为

$$f_K(a_1, a_2, \dots, a_m) = \bigwedge \{ \vee D_{ij} : D_{ij} \in M_0 \}$$

式中, $\vee D_{ij}$ 是 D_{ij} 中所有变量 a 的析取。

假设 $K=(U,A)$ 是一个区间值信息系统, 那么属性子集 $B \subseteq A$ 是 K 的一个分类约简, 当且仅当 $\bigwedge_{a_k \in B} a_k$ 是辨识函数 f_K 的极小析取范式^[19]。

根据表 1, 得到布尔函数:

$$f_K(a_1, a_2, a_3, a_4) = (a_2 \vee a_3) \wedge (a_3 \vee a_4) \wedge (a_2 \vee a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) \wedge a_3 \wedge (a_2 \vee a_4) \wedge (a_1 \vee a_2 \vee a_4)$$

经过简化计算(运用吸收律), 得到了布尔函数的极小析取范式:

$$f_K(a_1, a_2, a_3, a_4) = (a_2 \vee a_4) \wedge a_3 = (a_2 \wedge a_3) \vee (a_3 \wedge a_4)$$

这样系统有两个分类约简: $\{a_2, a_3\}$ 和 $\{a_3, a_4\}$ 。

同样, 根据表 2, 得到布尔函数:

$$f_K(a_1, a_2, a_3, a_4) = a_4 \wedge (a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) \wedge a_3 \wedge (a_2 \vee a_4) \wedge (a_1 \vee a_2 \vee a_4)$$

经过简化计算(运用吸收律), 得到了布尔函数的极小析取范式:

$$f_K(a_1, a_2, a_3, a_4) = a_3 \wedge a_4$$

这样系统仅有一个分类约简: $\{a_3, a_4\}$ 。

布尔函数的极小析取范式的每一个合取式中所构成的元素集合都是系统的分类约简。如果在布尔函数的合取范式的析取式中只取辨识矩阵中第 i 列(或第 i 行)的元素, 那么得到关于 u_i 类的可辨识函数, 用 f_i 表示:

$$f_i(a_1, a_2, \dots, a_m) = \bigwedge_{\{j: D_{ij} \in M_0\}} (\vee D_{ij})$$

则 f_i 的极小析取范式的每一个合取式中所构成的元素集合都是系统中 u_i 类的一个局部分类约简。每一个局部分类约简都包含了 u_i 类能从别的类中识别出来的类 u_i 的最小属性集合。

根据表 1, 得到 u_i 的布尔函数 $i=1, 2, \dots, 5$, 然后得到每个类的局部分类约简如下:

因为

$$f_1(a_1, a_2, a_3, a_4) = (a_2 \vee a_3) \wedge (a_3 \vee a_4) \wedge (a_2 \vee a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) = (a_2 \vee a_3) \wedge (a_3 \vee a_4) = (a_2 \wedge a_4) \vee a_3$$

所以 $\{a_3\}$ 和 $\{a_2, a_4\}$ 是 u_1 的分类约简, 也就是说 $re(u_1) = \{\{a_3\}, \{a_2, a_4\}\}$ 。

同样, 由 $f_2(a_1, a_2, a_3, a_4) = (a_2 \wedge a_4) \vee a_3$, 得

$$re(u_2) = \{\{a_3\}, \{a_2, a_4\}\}$$

由 $f_3(a_1, a_2, a_3, a_4) = (a_1 \wedge a_3) \vee (a_2 \wedge a_3) \vee (a_3 \wedge a_4)$, 得

$$re(u_3) = \{\{a_1, a_3\}, \{a_2, a_3\}, \{a_3, a_4\}\}$$

由 $f_4(a_1, a_2, a_3, a_4) = (a_2 \wedge a_3) \vee (a_3 \wedge a_4)$, 得

$$re(u_4) = \{\{a_2, a_3\}, \{a_3, a_4\}\}$$

由 $f_5(a_1, a_2, a_3, a_4) = a_2 \vee a_4$, 得

$$re(u_5) = \{\{a_2\}, \{a_4\}\}$$

同样, 根据表 2, 得到 u_i 的布尔函数 $i=12, 3, 4, 5$, 然后得到每个类的局部分类约简如下:

因为

$$f_{12}(a_1, a_2, a_3, a_4) = a_4 \wedge (a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) = a_4$$

所以 $\{a_4\}$ 是 u_{12} 的分类约简, 即

$$re(u_{12}) = \{\{a_4\}\}$$

同样, 由 $f_3(a_1, a_2, a_3, a_4) = a_3 \wedge a_4$, 得

$$re(u_3) = \{\{a_3, a_4\}\}$$

由 $f_4(a_1, a_2, a_3, a_4) = (a_2 \wedge a_3) \vee (a_3 \wedge a_4)$, 得

$$re(u_4) = \{\{a_2, a_3\}, \{a_3, a_4\}\}$$

由 $f_5(a_1, a_2, a_3, a_4) = a_2 \vee a_4$, 得

$$re(u_5) = \{\{a_2\}, \{a_4\}\}$$

2.3 分类规则

计算分类约简后, 可以发现隐藏在区间值信息系统中的分类知识, 并以分类规则的形式表述如下:

如果对于所有的 $a_k \in B$ 有 $a_k(x) \in [l_k^B, u_k^B]$, 那么样本 x 应该被分在类 u_i 。

在表 1 中, 根据每个类的分类约简, 可以发现隐藏在区间值信息系统中的所有分类规则, 表示如下:

r_1 : 如果 $a_3(x) \in [19, 30]$, 那么 $x \in u_1$ 。

r_1' : 如果 $a_2(x) \in [44, 68]$ 和 $a_4(x) \in [5, 16]$, 那么 $x \in u_1$ 。

r_2 : 如果 $a_3(x) \in [53, 60]$, 那么 $x \in u_2$ 。

r_2' : 如果 $a_2(x) \in [74, 80]$ 和 $a_4(x) \in [5, 19]$, 那么 $x \in u_2$ 。

r_3 : 如果 $a_1(x) \in [74, 97]$ 和 $a_3(x) \in [63, 107]$, 那么 $x \in u_3$ 。

r_3' : 如果 $a_2(x) \in [65, 116]$ 和 $a_3(x) \in [63, 107]$, 那么 $x \in u_3$ 。

r_3'' : 如果 $a_3(x) \in [63, 107]$ 和 $a_4(x) \in [56, 107]$, 那么 $x \in u_3$ 。

r_4 : 如果 $a_2(x) \in [73, 276]$ 和 $a_3(x) \in [108, 205]$, 那么 $x \in u_4$ 。

r_4' : 如果 $a_3(x) \in [108, 205]$ 和 $a_4(x) \in [104, 204]$, 那么 $x \in u_4$ 。

r_5 : 如果 $a_2(x) \in [38, 42]$, 那么 $x \in u_5$ 。

r_5' : 如果 $a_4(x) \in [29, 41]$, 那么 $x \in u_5$ 。

此外, 如果用系统分类约简 $B_1 = \{a_2, a_3\}$, 那么可以得到分类规则集 $r_1, r_2, r_3', r_4', r_5'$ 。同样, 如果用系统分类约简, $B_2 = \{a_3, a_4\}$, 那么可以得到分类规则集 r_1, r_2, r_3', r_4, r_5 。

在表 2 中, 根据每个类的分类约简, 可以发现隐藏在区间值信息系统中的所有分类规则, 表示如下:

r_1 : 如果 $a_4(x) \in [5, 18]$, 那么 $x \in u_{12}$ 。

r_2 : 如果 $a_3(x) \in [63, 107]$ 并且 $a_4(x) \in [56, 107]$, 那么 $x \in u_3$ 。

r_3 : 如果 $a_2(x) \in [73, 255]$ 和 $a_3(x) \in [108, 205]$, 那么 $x \in u_4$ 。

r_3' : 如果 $a_3(x) \in [108, 205]$ 和 $a_4(x) \in [104, 204]$, 那么 $x \in u_4$ 。

r_4 : 如果 $a_2(x) \in [38, 42]$, 那么 $x \in u_5$ 。

r_4' : 如果 $a_4(x) \in [29, 41]$, 那么 $x \in u_5$ 。

此外, 如果用系统分类约简, $B = \{a_3, a_4\}$, 那么可以得到分类规则集 r_1, r_2, r_3', r_4' 。

显然, 对不同粒度的信息系统进行分类, 从粒度粗的系统中提取出来的 if-then 规则少, 需要的计算量小。

3 分类精度

3.1 实际生活数据描述

表 5 和表 6 是带有地表物类别标记数据的训练样本和测

试样本的描述。数据是 2000 年 11 月 12 日在香港的西北部得到的 SPOT-4 多光谱数据。与文献[19]不同的是,这里已经将两类地表物水域和泥滩的合并标记成为一类了(即粒度

变粗了)。故合并标记后从训练样本得到的光谱波段反射灰度区间总体覆盖了合并前对应的灰度区间(见表 1 和表 2)。粒度变粗以后在这里所反映的是相应的灰度区间变大了。

表 5 训练样本描述

陆地 覆盖物	样本 数量	光谱波段															
		绿光(a ₁)				红光(a ₂)				近红外光谱(a ₃)				远红外光谱(a ₄)			
		均值	方差	最小值	最大值	均值	方差	最小值	最大值	均值	方差	最小值	最大值	均值	方差	最小值	最大值
水域和 泥滩	120	73.83	39.87	62	86	66.63	131.39	44	89	40.36	268.35	8	73	11.45	11.41	5	18
居民区	60	85.02	36.08	74	97	90.40	171.36	65	116	84.92	129.54	63	107	81.57	172.32	56	107
工业区	30	146.20	2300.79	51	242	174.73	2590.20	73	276	156.30	597.94	108	205	154.23	631.22	104	204
植被区	60	55.60	1.26	54	57	40.05	1.71	38	42	116.38	133.66	94	139	34.95	10.90	29	41

表 6 测试样本描述

陆地 覆盖物	样本 数量	光谱波段															
		绿光(a ₁)				红光(a ₂)				近红外光谱(a ₃)				远红外光谱(a ₄)			
		均值	方差	最小值	最大值	均值	方差	最小值	最大值	均值	方差	最小值	最大值	均值	方差	最小值	最大值
水域和 泥滩	66	70.65	62.17	55	71	62.42	228.86	33	62	51.06	634.33	1	51	30.12	607.95	0	30
居民区	33	82.88	46.61	70	96	82.91	72.27	66	99	78.73	42.83	66	91	75.36	29.49	65	86
工业区	30	135.47	2034.60	46	225	156.70	2648.01	54	259	139.07	1441.38	64	215	135.77	984.19	74	198
植被区	36	56.08	1.91	54	58	40.42	3.34	37	44	115.97	159.28	91	141	35.28	10.66	29	41

经统计计算后,得到对应地表物的光谱波段的反射灰度 $[l_i^k, u_i^k], i=1,2,3,4, k=1,2,3,4$, 其中

$$l_i^k = \text{int}(\max\{\mu_i^k - 2\sigma_i^k, 0\}) + 1$$

$$u_i^k = \text{int}(\min\{\mu_i^k + 2\sigma_i^k, 255\})$$

表 1 是根据文献[19]原始的数据集转化后得到的区间值信息系统,而表 2 是根据表 5(粒度变粗以后)的数据集转化后得到的区间值信息系统。

3.2 分类精度

当从样本集中导出分类规则集后,必须对分类规则的精度做出一些评估。标准做法第一步先构造出一个 $I \times (I+1)$ 的一个误差矩阵(error matrix),其中第 i 行、第 j 列的元素为 $n_{i,j}$,这里 $n_{i,j}, j=1,2,\dots,I$, 是第 i 类样本中被分到第 j 类的样本个数,而 $n_{i,I+1}$ 是第 i 类样本中不能被识别的个数。

下面定义几个用来评价分类质量的常用度量。

根据 Congalton 等^[25]的公式,第 j 类地表物的产家精度(producer's accuracy)定义为:

$$p_j = \frac{n_{j,j}}{\sum_{i=1}^I n_{i,j}}, j=1,2,\dots,I$$

第 i 类地表物的用户精度(user's accuracy)定义为:

$$q_i = \frac{n_{i,i}}{\sum_{j=1}^I n_{i,j}}, i=1,2,\dots,I$$

分类的总体精度(overall accuracy)定义为:

$$\alpha_o = \frac{\sum_{i=1}^I n_{i,i}}{\sum_{i=1}^I J_i}$$

式中, J_i 是第 i 类地表物的样本数。

用分类约简 $B=\{a_3, a_4\}$ 和 4 条规则 r_1, r_2, r_3', r_4' 得到经我们合并后的区间值信息系统的训练样本和测试样本的分类精度,见表 7 和表 8。我们发现,与水域和泥滩合并前相比较(见文献[19]),测试样本的产家精度、用户精度和总体分类精度都比合并前下降了,其原因是类别的粒度标记变粗了。

表 7 用训练样本的分类约简 $B=\{a_3, a_4\}$ 和 4 条规则 r_1, r_2, r_3', r_4' 的分类精度

训练样本	水域和 泥滩	居民区	工业区	植被区	不可辨 识区	用户精度
水域和泥滩 (120 个样本)	115	0	0	0	5	115/120=0.958
居民区 (60 个样本)	0	58	2	0	0	58/60=0.917
工业区 (30 个样本)	0	0	28	0	2	28/30=0.933
植被区 (60 个样本)	0	0	0	59	1	59/60=0.983
产家精度	115/115 =1.0	58/58 =1.0	28/30 =0.933	59/59=1.0	5/5=1.0	总体精度=260/270 =0.963

表 8 用测试样本的分类约简 $B=\{a_3, a_4\}$ 和 4 条规则 r_1, r_2, r_3', r_4' 的分类精度

训练样本	水域和 泥滩	居民区	工业区	植被区	不可辨 识区	用户精度
水域和泥滩 (66 个样本)	34	12	0	8	12	34/66=0.515
居民区 (30 个样本)	3	14	1	1	11	14/30=0.467
工业区 (30 个样本)	4	9	6	0	11	6/30=0.2
植被区 (36 个样本)	5	3	0	15	13	15/36=0.417
产家精度	34/46=0.739	14/38=0.368	6/7=0.857	15/24=0.625	12/24=0.5	总体精度=69/162=0.426

结束语 本文提出了基于粗糙集的空间遥感数据的多粒度标记分类方法,并给出了相应粒度下的分类精度。根据实际数据我们认为,分类标记的粒度越粗,规则集越少,规则泛化能力强,但相应的分类精度一般会降低。尽管粒度大的信息系统比粒度小的需要的计算少,分类精度会降低一些,但是在对分类的精度要求不高的时候,依然可以采用粒度大的信息系统以便于节省时间和精力。

参 考 文 献

[1] Pawlak Z. Rough sets [J]. International Journal of Information

- and Computer Sciences, 1982, 11, 341-356
- [2] 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
 - [3] Ahlqvist O, Keukelaar J, Oukbir K. Rough classification and accuracy assessment [J]. *International Journal Geographical Information Science*, 2000, 14: 475-496
 - [4] Ahlqvist O, Keukelaar J, Oukbir K. Rough and fuzzy geographical data integration [J]. *International Journal Geographical Information Science*, 2003, 17: 223-234
 - [5] Aldridge C H. A theory of empirical spatial knowledge supporting rough set based knowledge discovery in geographic databases[D]. New Zealand: University of Otago, 1998
 - [6] Bittner T, Stell J. Stratified rough sets and vagueness [C]//Kuhn W, Worboys M F, Timpf S, eds. *Spatial Information Theory. Proceedings of International Conference on Foundations of Geographic Information Science (COSIT 2003)*. Karlsruhe Ittingen, Switzerland, September 2003, 2825: 270-286
 - [7] Wang S L, Li D, Shi W Z, et al. Geo-rough space [J]. *Geo-Spatial Information Science*, 2002, 6: 54-61
 - [8] Worboys M F. Computation with imprecise geographical data [J]. *Computers Environment, Urban Systems*, 1998, 22: 85-106
 - [9] Bardossy A, Samaniego L. Fuzzy rule based classification of remotely sensed images [J]. *IEEE Transaction on Geoscience, Remote Sensing*, 2002, 40: 362-374
 - [10] Benediktsson J A, Swain P H, Erson O K. Neural network approaches versus statistical methods in classification of multi-source remote sensing data [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1990, 28: 540-552
 - [11] Ji M. Using fuzzy sets to improve cluster labelling in unsupervised classification [J]. *International Journal of Remote Sensing*, 2003, 24: 657-671
 - [12] Kavzoglu T, Mather P M. The use of backpropagating artificial neural networks in land cover classification [J]. *International Journal of Remote Sensing*, 2003, 24: 4907-4938
 - [13] Leung Y, Leung K S. An intelligent expert system shell for knowledge-based geographic information systems: 1. the tools [J]. *International Journal of Geographical Information Systems*, 1993, 7: 189-199
 - [14] Leung Y, Leung K S. An intelligent expert system shell for knowledge-based geographic information systems: 2. some applications [J]. *International Journal of Geographical Information Systems*, 1993, 7: 201-213
 - [15] Paola J D, Schowengerdt R A. A detailed comparison of back-propagation neural network, maximum likelihood classifiers for urban land use classification [J]. *IEEE Transactions on Geoscience, Remote Sensing*, 1995, 33: 981-996
 - [16] Wang F. Improving remote sensing image analysis using fuzzy information representation [J]. *Photogrammetric Engineering, Remote Sensing*, 1990, 56: 1163-1169
 - [17] Wang S L, Wang X Z, Shi W Z. Development of a data mining method for land control [J]. *Geo-Spatial Information Science*, 2001, 4: 68-76
 - [18] Worboys M F. Imprecision in finite resolution spatial data [J]. *GeoInformatica*, 1998, 2: 257-279
 - [19] Leung Y, Fung T, Mi J-S, et al. A rough set approach to the discovery of classification rules in spatial data [J]. *International Journal of Geographical Information Science*, 2007, 21: 1033-1058
 - [20] Leung Y, Manfred M F, Wu W-Z, et al. A rough set approach for the discovery of classification rules in interval-valued information systems [J]. *International Journal of Approximate Reasoning*, 2008, 47: 233-246
 - [21] Yao J T. Recent developments in granular computing: A bibliometrics study [C]//*Proceedings of IEEE International Conference on Granular Computing*. Hangzhou, China, Aug. 2008: 74-79
 - [22] 苗夺谦, 王国胤, 刘清, 等. 粒计算: 过去、将来与展望[M]. 北京: 科学出版社, 2007
 - [23] 张铃, 张拔. 问题求解理论及其在应用: 商空间粒度计算理论及其在应用[M]. 北京: 清华大学出版社, 2007
 - [24] Wu W-Z, Leung Y. Theory and applications of granular labelled partitions in multi-scale decision tables [J]. *Information Sciences*, 2011, 181: 3878-3897
 - [25] Congalton R G, Green K. Assessing the accuracy of remotely sensed data: principles and practices [M]. Boca Raton: Lewis Publications, 1999

(上接第 22 页)

- [95] Ji Q, Marefat M. Extracting and identifying form features: a Bayesian approach [J]. *Computer-Aided Design*, 1995, 27 (6): 435-454
- [96] Vandenbrande J, Requicha A. Spatial reasoning for the automatic recognition of machinable features in solid models [J]. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1993, 15(12): 1269-1285
- [97] Zhou X H, Qiu Y J, Hua G R, et al. A feasible approach to the integration of CAD and CAPP [J]. *Computer-Aided Design*, 2007, 39(4): 324-348
- [98] Kalyanapathy V, Lin E, Minis I. Group technology code generation over the internet ETC97/CIE-4290 [C]//*Proceedings of ASME Design Engineering Technique Conferences*. Sacramento, CA, 1997: 14-17
- [99] Hermann J W, Singh G. Design similarity measures for process planning and design evaluation [R]. TR97-74. University of Maryland, 1997
- [100] Iyer S, Nagi R. Identification and ranking of similar parts in agile manufacturing [C]//*Proceedings of the 4th Industrial Engineering Research Conference*, 1995
- [101] 崔晨阳, 石教英. 三维模型检索中的特征提取技术综述 [J]. *计算机辅助设计与图形学学报*, 2004, 16(7): 882-889
- [102] Shilane P, Min P, Kazhdan M, et al. The princeton shape benchmark [C]//*Proceedings of the Shape Modeling Int'l 2004*. Los Alamitos, 2004: 167-178
- [103] <http://shapes.aim-at-shape.net/>
- [104] Jayanti S, Kalyanaraman Y, Iyer N, et al. Developing an engineering shape benchmark for CAD models [J]. *Computer-Aided Design*, 2006, 38(9): 939-953
- [105] <http://www.designrepository.org/datasets/>
- [106] 高曙明, 何发智. 异构 CAD 系统集成技术综述 [J]. *计算机辅助设计与图形学学报*, 2009, 21(5): 561-568
- [107] 陶松桥, 王书亭, 郑坛光, 等. 基于非精确图匹配的 CAD 模型检索方法 [J]. *计算机辅助设计与图形学学报*, 2010, 22(3): 545-552
- [108] 开兴, 张树生, 白晓亮, 等. 一种属性图同构的三维 CAD 模型局部匹配算法 [J]. *西安交通大学学报*, 2010, 44(11): 56-60
- [109] 刘玉生, 蒋玉芹, 高曙明. 模型驱动的复杂产品系统设计建模综述 [J]. *中国机械工程*, 2010, 21(6): 741-749