基于 GOODALL 相近指数的遥感图像和其它空间数据综合分类方法*

田 青
(中国科学院遥感应用研究所 北京 100101)

Enrico Feoli
(Department of Biology Trieste University Italy)

摘 要 介绍 David W. Goodall 的基于概率的相近指数理论,研究它被应用在遥感图像和其它空间数据综合分类中的可能性,并首次在 GRASS 环境下实现了基于 David W. Goodall 的相近指数的遥感图像和其它空间数据综合分类算法,并对该算法进行了测试,将分类结果与其它几种较流行的分类方法结果进行了比较。

关键词 遥感图像,空间数据,综合分类算法,相近指数

至今已经有很多分类算法被应用在遥感图像分类中,譬如常用的最大似然分类。但是开发能够同时处理遥感图像和其它空间数据的计算机分类算法仍然是一个非常活跃的研究领域,这被称做多源数据集成[1]。作者研究了 David W. Goodall 的基于概率的相近指数理论[2]和它被应用在遥感图像和其它空间数据综合分类中的可能性,并在 GRASS 环境下实现了基于 David W. Goodall 的相近指数的遥感图像和其它空间数据综合分类算法。还将分类结果同目前较流行的最大似然分类和基于证据分类结果进行了比较。试图通过比较这些不同的分类结果回答如下问题:(1)基于 David W. Goodall 的相近指数的遥感图像和其它空间数据综合分类算法的效果。(2)其它空间数据能在何种程度上辅助遥感图像提高土地覆盖和利用分类的精度。

1 David W. Goodall 的基于概率的相近指数理论

早在 60 年代,David W. Goodall 就提出了基于概率的相近指数(Affinity Index),用于描述一个个体与一个已知类之间的相近程度。David W. Goodal 的相近指数与众不同之处在于它并不直接考虑个体属性值本身而是考虑一组个体中的每一个与一个已知类之间的相对亲近关系。因此它能够处理多种不同种类的属性:数值型属性(quantitative attribute),定性化属性(qualitative attribute)及排序型属性(ranked attribute)。

在数字分类学领域里,经常会发生这样的情形:一组紧密相关的个体组成的类已被识别出来,而人们希望知道另一个个体是否可以被恰当地归入这个已知类中。或着判别一个个体应被归入几个已知类中的哪一个。在判别过程中,人们需要比较具有特定属性的个体可能发生在每个不同类中的可能性。这个问题也可以表达为估计一组未被分类的个体中的每一个个体事实上接近一个特定类的可能性。

David W. Goodall 的相近指数是针对数值型属性,定性化属性和排序型属性分别进行定义的。对每种属性一个标准值(norm)被定义来代表一个类。然后所有可能的其它值根据它们与标准值的相似程度进行排序。每个个体在某属性上对该特定类的相近指数通过下面分别对不同类型的属性定义的方法进行计算。最后,一个个体与一个已知类间的相近指数可以通过综合它在这些不同属性上的相近指数得到。

1.1 定性型属性

取标准值为众数,其它所有值与标准值的相似

* 此工作由 ICS UNIDO (International Center for Science and High Technology-United Nations Industrial Development Organization)资助完成。基于证据分类由 Trieste 大学生物系 Cristina Milesi 完成,在此表示感谢。欢迎对此感兴趣的人从 GRASS 的 FTP 地址获取该计算机程序(Quality)及相关使用说明,并对此加以检验。

收稿日期:1998-04-02;收到修改稿日期:1999-03-26

程度取决于它们在已知类中发生的频率。如果我们用 A_{ij}来表示第 i 个属性(共有 r_i 个不同值)上,具有值 j 的个体在第 i 个属性上与某个特定类的相近程度,n_{ij}是值j 在该类中发生的次数,A_{ik}表示第 i 个属性上具有值k 的个体在第 i 个属性上与某个特定点的相近程度,n_{ik}是值 k 在该类中发生的次数。那么有,

$$n_{ij} < n_{ik} \supset \mathrm{A}_{ij} < A_{ik} \tag{1}$$

$$n_{ij} = n_{ik} \supset \mathrm{A}_{ij} = A_{ik} \tag{2}$$

如果用 m_{ij}表示在将要分类的集合中第 i 个属性具有值j 的个体数目,s 是所有比值j 在第 i 个属性上与该类更相近的值的集合,那么这些个体在该属性上拥有的与该类的相似程度的最小可能性 p_{ij},即相近指数,可估算为:

$$p_{ij} = \sum_{k\in \mathrm{S}} m_{ik} \Big/ \sum_{k=1}^{r_i} m_{ik}$$

$$S = \{k : A_{ik} \geqslant \mathrm{A}_{ij}\} \tag{3}$$

1.2 排序型属性

取标准值为中值,在标准值的相同方向上的两个个体,即两个个体都大于或都小于标准值,与标准值更近的更相似。在标准值的相反方向上的两个个体,即两个个体中的一个大于标准值而另一个小于标准值,它们与标准值的相似程度由它们的《尾》集合中个体数目决定。如果某类的标准值为1,值 j 的《尾》集合是这样定义的:如果 j 小于或等于 1,则 j 的《尾》集合由所有属于该类的小于或等于 j 的个体组成;如果 j 大于 1,则由所有属于该类的大于 j 的个体组成。因此《尾》集合中个体数目 t_{ij}可表示为:

$$t_{ij} = \sum_{k=1}^{j} n_{ik},\ j \leqslant 1 \tag{4}$$

或

$$t_{ij} = \sum_{k=j}^{r_i} n_{ik},\ j > 1 \tag{5}$$

然后有,

$$t_{ij} = t_{ik} \supset \mathrm{A}_{ij} = A_{ik} \tag{6}$$

$$t_{ij} < t_{ik} \supset \mathrm{A}_{ij} < A_{ik} \tag{7}$$

1.3 数值型属性

取标准值为平均值,与标准值差别小的个体与标准值更相似,与标准值差别相同的两个个体,它们与标准值的相似程度由它们的《尾》集合中个体数目决定。

如果用 μ_i 表示平均值,x_{ij}、x_{ik}是两个不同个体在第 i 个属性上的值,那么有

$$|x_{ij} - \mu_i| < |x_{ik} - \mu_i| \supset \mathrm{A}_{ij} > A_{ik} \tag{8}$$

$$(|x_{ij} - \mu_i| = |x_{ik} - \mu_i|) \wedge (\mathrm{t}_{ij} < t_{ik}) \supset \mathrm{A}_{ij} < A_{ik} \tag{9}$$

$$(|x_{ij} - \mu_i| = |x_{ik} - \mu_i|) \wedge (\mathrm{t}_{ij} = t_{ik}) \supset \mathrm{A}_{ij} = A_{ik} \tag{10}$$

则在第 i 个属性上的具有值 x_{ij}的个体,在该属性上与该类的相近程度可以由(3)得到。

当 p_{ij}对 α 个属性分别求得后,综合相近指数 p_k 可用如下方法求得:

$$p_k = -2\sum_{i=1}^{\alpha} \ln p_{ij} \tag{11}$$

这是一个自由度为 2α 的 χ^2 分布[3]。

2 David W. Goodall 的相近指数理论的应用

David W. Goodall 的基于概率的相近指数理论虽然早在 60 年代就已提出,并且他在以后的数十年中不断修改和完善,建立了一整套基于概率的统计分析理论。这些方法曾一度被应用于生物生态学等领域。但或许是因为他的思想颇为费解,没有很多人真正下工夫去理解,至今还没有人尝试过将它应用在遥感领域中。

David W. Goodall 的相近指数可被应用在遥感图像和其它空间数据的监督分类中。代表不同地物的一组训练区被确定后,计算每个像素与各个训练区之间的相近指数,比较某个像素与各个训练区间的相近指数,进而该像素被归入相近指数最大的一类。在这里,遥感图像可被看作数值型数据,地质分类图可视为定性化数据,而坡度分类图为排序型数据。

下面举例说明如何利用 David W. Goodall 的相近指数对遥感图像和其它空间数据进行监督分类。假设我们要对某个地区进行土地利用分类,已有数据为TM 图像(200 * 200)和已与图像配准的土壤类型图。为简便说明,这里只选用 TM 图像的一个波段。土壤类型包括 4 种,1、2、3、4 分别表示褐土、潮土、水稻土和沙土。经实地考察,在图上选出小麦的训练区,表 1 中对小麦训练区中所有像素在 TM 图像和土壤类型两个属性上不同值发生频率进行了统

计。

表 1　被识别出的小麦类

Table 1　The recognized wheat cluster

属性	值	发生频率
TM 图像	160	1
	161	3
	162	6
	163	20
	164	30
	165	40
	166	20
	167	8
	168	6
	169	6
	170	4
	171	4
	173	2
	其它	0
土壤类型	1	120
	2	30
	其它	0

第一个属性 TM 图像，可认为是数值型属性，被识别出的小麦类的标准值(norm)＝(160＊1＋161＊3＋162＊6＋163＊20＋164＊30＋165＊40＋166＊20＋167＊8＋168＊6＋169＊6＋170＊4＋171＊4＋173＊2)/150＝165.22，而将要被分类的像素数目为 200＊200＝40000。

在表 2 中，所有出现在将要被分类的图像中的像素的不同值按照它们与标准值的相似程度进行了排序，并计算了它们与已知小麦类的相近指数。

表 2　在第一个属性 TM 图像上的相近指数

Table 2　Affinity index to the first attrtibute：TM image

值	与标准值的差别	发生频率	相近指数
0	|0－165.22|＝165.22	10	10/40000＝0.00025
1	|1－165.22|＝164.22	20	30/40000＝0.00075
2	|2－186.22|＝163.22	20	50/40000＝0.00125
…	…	…	…
168	|168－165.22|＝2.78	600	36600/40000＝0.915
163	|163－165.22|＝2.22	400	37000/40000＝0.925
167	|167－165.22|＝1.78	1000	38000/40000＝0.95
164	|164－165.22|＝1.22	1000	39000/40000＝0.975
166	|166－165.22|＝0.78	500	39500/40000＝0.9875
165	|165－165.22|＝0.22	500	40000/40000＝1.0

第二个属性土壤类型，可认为是定性化属性，每个不同土壤类型与已知小麦类的相近指数被列在表 3 中。

表 3　在第二个属性土壤类型上的相近指数

Table 3　Affinity index to the second attribute：soil type

值	在已知小麦类中发生频率	在整个土壤类型图中发生频率	相近指数
4	0	5000	(5000＋10000)/40000＝0.375
3	0	10000	(5000＋10000)/40000＝0.375
2	30	5000	(5000＋10000＋5000)/40000＝0.50
1	120	20000	(5000＋10000＋5000＋20000)/40000＝1.0

因而图上图像灰度值为 163，土壤类型为 1 的某点，与已知小麦类的综合相近指数为：－2＊(ln0.925＋ln1.0)，而图像灰度值为 2，土壤类型为 4 的某点与已知小麦类的综合相近指数为：－2＊(ln0.00125＋ln0.375)。依此，可求出图上各点与已知小麦类的综合相近指数。如果我们在图上再选出其它土地利用类型(如水稻等)的训练区，就可用同样的方法，求出图上各点与这些已知类的综合相近指数。比较某点与所有这些已知类的综合相近指数，可把它归入综合相近指数最大的一类。

基于 David W. Goodall 的相近指数，我们在 GRASS 环境中，实现了一个对遥感图像和其它空间数据进行综合分类的计算机程度(r.affinity)，并对该算法进行了测试，将分类结果与其它几种较流行的分类方法进行了比较。

3　实验区，实验数据及分类结果比较

意大利北部的 TARVISIO 森林地区被选用来测试我们的算法。在该森林区中能够发现多种植被：如草地，松等及居民区。用于分类的遥感数据为 TM(第 3,4,5 波段)，其它辅助空间数据有：地质图，DEM，坡向分类图及坡度分类图。这些辅助空间数据被选用，是因为在山区植被种类与地质结构，高度，坡向和坡度紧密相关，相信这些辅助空间数据的使用会提高分类精度。

经过实地考察，代表 11 类不同地物的训练样本被选出用于对整个山区进行分类。它们是：草地，居民区，阴影，松，云杉，混有山毛榉的云杉，山毛榉，水体，岩石，牧草和混合林木。我们用相同的训练区进

行了5种分类:最大似然分类(只用遥感数据),David W. Goodall 的相近指数分类(只用遥感数据),David W. Goodall 的相近指数分类(遥感数据和其它辅助空间数据并用),基于证据分类(只用遥感数据),基于证据分类(遥感数据和其它辅助空间数据并用)。见图版Ⅰ图1—图5。最大似然分类用 GRASS 完成,而基于证据分类是由 IDRISI 完成的。

另一组随机选择的不同于用于分类训练区的样本被用来测试分类结果。分类后,我们对用于测试的样本计算了差错矩阵[4],以反映分类精度。综合分类精度=被正确划分的检测样本数目/所有检测样本的数目*100。算法精度=各类中被正确划分的检测样本数目/该类的检测样本数目*100,它反应了某个特定类的检测样本被正确分类的程度。用户精度=各类中被正确划分的检测样本数目/被划分为该类的检测样本的数目*100,它说明了划分到某个特定类中的样本真正能代表地面上该类的可能性。

在下面的差错矩阵中(表4—表8),不同数字分别代表以下类。

1:草地 2:居民区 3:阴影
4:松 5:云杉 6:混有山毛榉的云杉
7:山毛榉 8:水体 9:岩石
10:牧草 11:混合林木

差错矩阵

表4 只用遥感数据的最大似然分类的差错矩阵

Table 4 The confusion error matrix for Maximum-likelihood classification

	分类数据											
	类别号	1	2	3	4	5	6	7	8	9	10	11	用户精度/%
	1	93	54	0	0	0	0	1	28	0	5	1	51.10
	2	0	56	0	0	0	0	3	20	10	2	1	60.87
检测	3	0	0	181	0	0	0	0	3	0	0	0	98.37
	4	0	2	0	54	4	6	12	0	0	0	20	55.10
	5	0	0	0	52	35	0	0	0	0	0	0	40.23
数据	6	2	2	0	48	69	23	0	0	0	0	3	15.65
	7	0	2	0	7	0	3	120	0	0	0	42	68.97
	8	0	2	73	0	0	0	0	10	0	0	0	11.76
	9	0	1	0	0	0	0	0	3	253	0	0	98.44
	10	9	0	0	0	0	0	4	0	1	41	1	73.21
	11	3	0	0	20	38	13	0	0	0	1	15	16.67
算法精度/%		86.92	47.06	71.26	29.83	23.97	51.11	85.71	15.62	95.83	83.67	18.07	

综合分类精度=60.67%

表5 只用遥感数据的基于 Goodall 的相近指数分类的差错矩阵

Table 5 The confusion error matrix for classification based on Goodall's affinity index with only images

	分类数据											
	类别号	1	2	3	4	5	6	7	8	9	10	11	用户精度/%
	1	37	68	0	1	0	2	7	0	64	3	0	20.33
	2	5	28	0	0	0	0	10	0	37	11	1	30.43
检测	3	0	0	184	0	0	0	0	0	0	0	0	100.00
	4	1	1	0	7	2	33	11	0	0	1	42	7.14
	5	0	0	0	15	57	6	0	9	0	0	0	65.52
数据	6	1	4	0	41	42	35	2	4	0	2	16	23.81
	7	23	16	0	0	0	3	85	0	0	12	35	48.85
	8	0	0	83	1	0	0	1	0	0	0	0	0.00
	9	0	2	0	0	0	0	1	0	254	0	0	98.83
	10	17	0	0	0	0	0	2	0	2	35	0	62.50
	11	2	0	0	13	12	26	4	2	1	7	23	25.56
算法精度/%		43.02	23.53	68.91	8.97	50.44	33.33	69.11	0.00	70.95	49.30	19.66	

综合分类精度=51.31%

表 6　遥感数据和其它辅助空间数据并用的基于 Goodall 的相近指数分类的差错矩阵

Table 6　The confusion error matrix for classification based on Goodall's affinity index with images and other data

	分类数据												
	类别号	1	2	3	4	5	6	7	8	9	10	11	用户精度/%
	1	91	55	2	0	9	2	0	0	12	7	4	50.00
	2	22	33	0	0	0	0	3	0	3	31	0	35.87
检测	3	0	0	184	0	0	0	0	0	0	0	0	100.00
	4	5	18	0	57	0	3	7	0	0	0	8	58.16
	5	0	0	0	8	47	30	0	2	0	0	0	54.02
数据	6	0	0	0	2	28	106	0	10	0	0	1	72.11
	7	12	4	0	4	0	3	117	0	0	12	22	67.24
	8	0	0	7	0	0	0	0	78	0	0	0	91.76
	9	0	0	0	0	1	1	12	0	239	2	2	93.00
	10	1	0	0	0	0	0	0	0	0	55	0	98.21
	11	11	0	0	5	18	31	0	0	0	1	24	26.67
算法精度/%		64.08	30.00	95.34	75.00	45.63	60.23	84.17	86.67	94.09	50.93	39.34	

综合分类精度＝71.01%

表 7　只用遥感数据的基于证据分类的差错矩阵

Table 7　The confusion error matrix for classification based on evidential theory with only images

	分类数据													
	类别号	不能确定	1	2	3	4	5	6	7	8	9	10	11	用户精度/%
	1	1	85	82	0	0	0	0	0	0	0	12	2	46.70
	2	0	0	73	0	0	0	0	6	0	4	0	9	79.35
检测	3	0	0	0	69	0	0	0	0	115	0	0	0	37.50
	4	0	1	2	0	5	1	27	6	0	0	0	56	5.10
	5	0	0	0	0	22	55	10	0	0	0	0	0	63.22
数据	6	0	4	2	0	14	47	78	0	0	0	0	2	53.06
	7	0	0	9	0	0	0	22	95	0	0	0	48	54.60
	8	0	0	7	5	0	0	1	0	72	0	0	0	84.71
	9	6	0	10	0	0	0	0	0	0	240	0	1	93.39
	10	0	18	0	0	0	0	0	1	0	0	32	5	57.14
	11	2	3	0	0	0	16	55	0	0	0	0	14	15.56
算法精度/%			76.58	39.46	93.24	12.20	46.22	40.41	87.96	38.50	98.36	72.73	10.22	

综合分类精度＝56.34%

表 8　遥感数据和其它辅助空间数据并用的基于证据分类的差错矩阵

Table 8　The confusion error matrix for classification based on evidential theory with images and other data

	分类数据													
	类别号	不能确定	1	2	3	4	5	6	7	8	9	10	11	用户精度/%
	1	81	92	0	0	4	4	1	0	0	0	0	0	50.55
	2	64	4	19	0	1	3	1	0	0	0	0	0	20.65
检测	3	25	0	0	118	0	0	41	0	0	0	0	0	64.13
	4	0	0	1	0	83	0	14	0	0	0	0	0	84.69
	5	0	0	0	0	0	34	53	0	0	0	0	0	39.08
数据	6	0	2	0	0	3	43	99	0	0	0	0	0	67.35
	7	0	1	0	0	0	0	65	97	0	0	0	11	55.75
	8	6	0	0	2	0	1	0	0	76	0	0	0	89.41
	9	6	0	0	0	0	0	0	0	0	251	0	0	97.67
	10	0	49	0	0	0	0	0	0	0	0	7	0	12.50
	11	8	9	0	0	0	16	57	0	0	0	0	0	0.00
算法精度/%			58.60	95.00	98.33	91.21	33.66	29.91	100.00	100.00	100.00	100.00	0.00	

综合分类精度＝60.33%

通过比较这些分类结果，可以得出如下结论：

(1)采用了其它辅助空间数据的分类，其精度都大大高于那些只使用遥感数据的分类；

(2)在只用遥感数据的所有分类中，最大似然分类效果最好；

(3)当其它辅助空间数据也被用于基于 David W. Goodall 的相近指数分类时，其分类精度由 51.31%提高到 71.01%；

(4)当其它辅助空间数据也被用于基于证据的分类时,其分类精度由56.43%提高到60.33%;

(5)在对多源空间数据进行分类时,基于David W. Goodall的相近指数分类比基于证据的分类效果好;

(6)最好的分类结果(71.01%)由基于David W. Goodall的相近指数分类获得。

也就是说,其它空间数据确实能帮助提高遥感图像的分类精度,而基于Goodall的相近指数分类也的确是一种利用其他空间数据提高遥感图像分类精度的好方法。

参 考 文 献 (References)

1 Ncgia. The research plan of the National Center for Geographical Information and Analysis. *Int. J. Geographical Information Systern*, 1989, **3**(2): 117—136.

2 Goodall, D.W. Affinity Between an Individual and a Cluster in Numerical Taxonomy. BIOMETRIE. PRAXIMETRIE, 1968, IX, I: 53—55.

3 Goodall, D.W. A new similarity index on probability. Biometrics, 1966, **22**: 882—907.

4 Lillesand, T.M., Kiefer, R.W. Remote Sensing and Image Interpretation. 1994.

作 者 简 介

田青,1990年毕业于北京大学计算机系,获理学士学位。1993年于中国科学院遥感应用研究所获遥感与制图硕士学位。毕业后一直从事地理信息系统与遥感方面的研究工作。

An Algorithm for Spatial Data Integrated Classification Based on GOODALL's Affinity Index

TIAN Qing

(*Institute of Remote Sensing Applications, Chinese Academy of Sciences* 100101)

Enrico Feoli

(*Biology Department, Trieste University, Italy*)

Abstract Today many methods have been used in classifying remote sensing images. However, developing classification algorithm which is capable of processing both images and other ancillary spatial data still remains to be an active research area. In this paper, the affinity index of David W. Goodall based on probability was explained, and its application possibility in remote sensing and other spatial data integrated classification was studied. Based on Goodall's affinity index, a computer program for classifying both remote sensing and other spatial data was developed within GRASS environment. To see the result of this program, it was tested in a case study area and compared with other popular classification methods such as maximum likelihood classification and evidential classification. Through this study, we would like to know how the other spatial data can help improve the remote sensing image classification and whether the algorithm based on Goodall's affinity index is good in classifying remote sensing images and other ancillary spatial data in an integrate way.

Key words Remote sensing image, Spatial data, Multisource data classification, Affinity index.

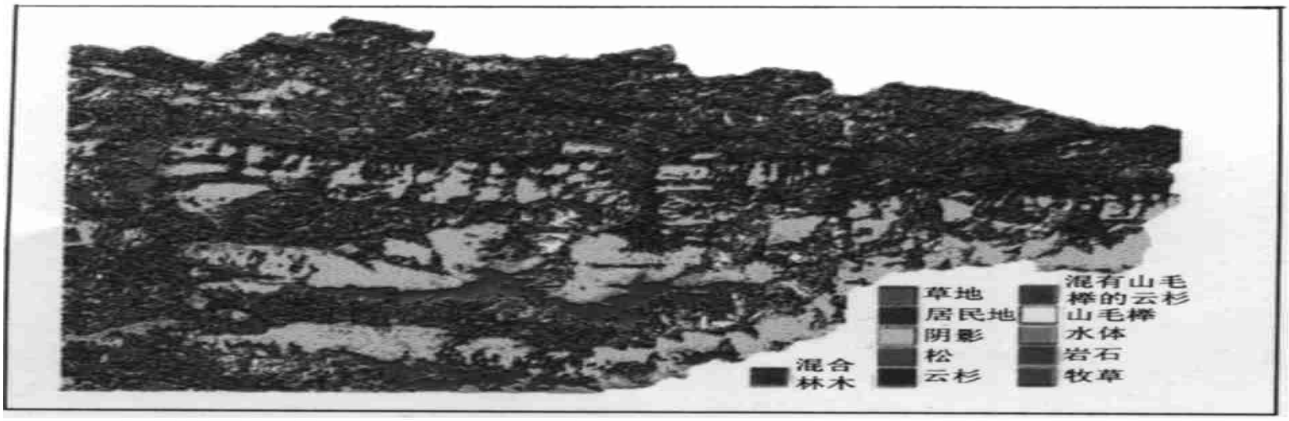

图 1　只用遥感数据的最大似然分类图

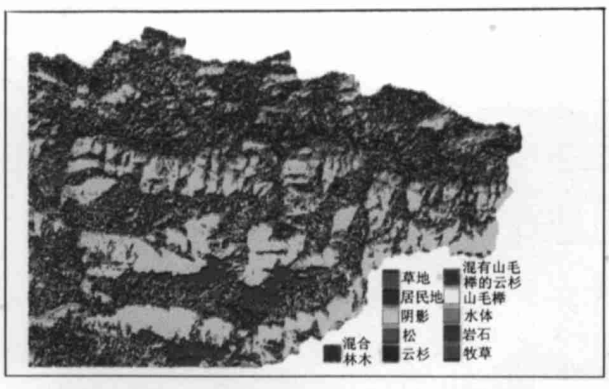

图 2　只用遥感数据的基于 Goodall 的相近指数分类图

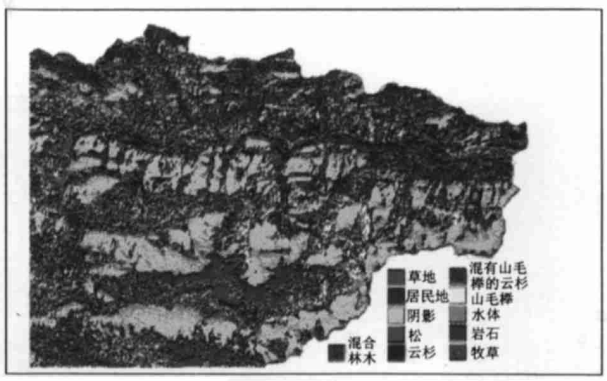

图 3　遥感数据和其它数据并用的基于 Goodall 的相近指数分类图

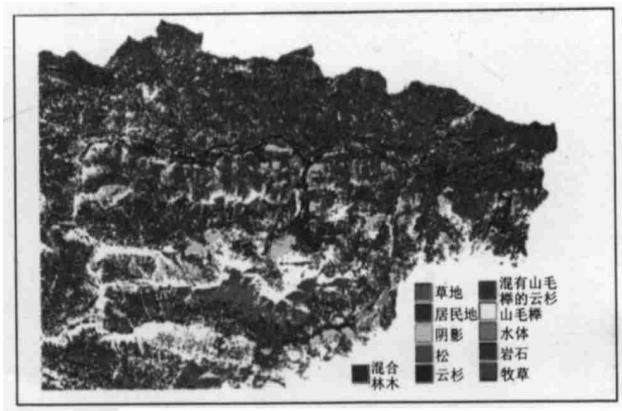

图 4　只用遥感数据的基于证据分类图

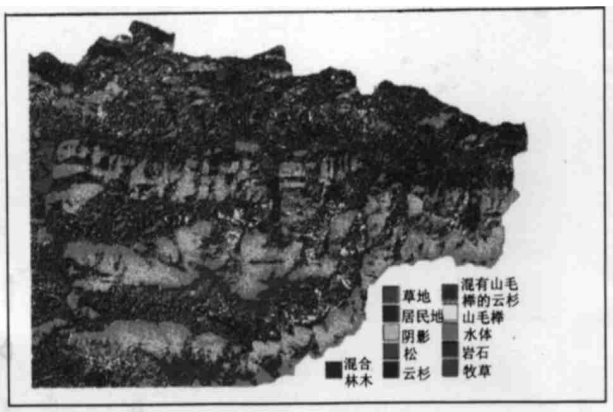

图 5　遥感数据和其它辅助空间数据并用的基于证据分类图