前期土地覆被数据辅助下的分类样本自动选取

刘 锟[1,2],杨晓梅[1*],张 涛[1,2]
(1. 中国科学院地理科学与资源研究所,北京 100101; 2. 中国科学院大学,北京 100049)

摘要: 将地学知识与影像标定相结合,一直是目视解译或计算机自动分类制图的主要手段。传统的目视解译方法能够充分利用地学知识,但需要大量的人力、物力,效率较低;计算机分类中尚未出现比较成熟的高效运用地学知识的分类方法。已有研究表明,分类样本可以作为地学知识的载体,将地学知识融入分类过程中;此外,无监督聚类可以显著提高样本选取的效率,有助于提供足够的样本,为将地学知识高效地融入计算机分类提供了一定的基础。本文提出一种以前期土地利用数据辅助与影像聚类相结合的样本自动选取方法。利用自动选取的样本,通过最大似然分类器对 TM 影像进行分类,并与手动选取样本分类的方法进行了对比分析。研究结果表明,在分类效果上,本文提出的前期土地覆被辅助下的分类样本自动选取方法,优于手动选取样本的方法,提高了分类效率。在水体、林地、园地、城镇建设用地等 7 种类型上的分类整体精度达到 84.18%,kappa 系数为 0.8066;手动选取样本进行分类的整体精度为 77.04%,kappa 系数为 0.7196。

关键词: 分类;样本;自动选取;LUCC
DOI: 10.3724/SP.J.1047.2012.00507

1 引言

目前,遥感制图中,地物的影像识别分类之研究重点:(1)地学知识的高效融入:即目视解译所依靠的专家知识,以及建立判别模型或函数时融入的地学理解,如术洪磊、毛赞猷利用 TM 影像,通过构造判别规则,对广东东莞市进行了土地利用/土地覆被分类,精度优于最大似然法[1]。在制定分区及影像选取的依据时,通过植被 NDVI 时间序列的特点来选取分类所用的影像时相[2]。(2)对影像的快速标定,包括对于样本的标定和直接对类别的标定。其中样本的标定即样本选取,其工作量大、效率低,且较容易受人为误差干扰[3-5]。而直接标定类别主要是对无监督聚类后的聚簇的标定,面临着与样本选取相同的困难。

因此,影像分类的一个重要发展方向是将地学知识有效地融入到影像的快速标定中。由于已有的土地利用/土地覆被的成果,是前人在融入了大量的地学知识的基础上分析提取出来的,所以,将现有的土地利用/土地覆被数据以一定的方式融入新的土地利用/土地覆被研究中去,成为了学者们首先想到的方法[2],大量的研究围绕着如何利用包括土地利用/土地覆被数据在内的地理信息系统(Geographical Information System ,GIS)数据构建知识库,直接对影像进行分类。但其流程复杂,精度受到地物分布趋势的限制,且判别规则的参数适用性受到不同地域的局限[4,6-8]。如何在自动识别中,将现有 GIS 数据直接高效地用于影像分类,尚未有令人满意的研究成果。

高质量的样本对于提高分类精度有着重要作用,因此,样本的选取成为了地学知识融入的极佳切入点。有学者将 GIS 数据用于样本的标定,在一定程度上利用了以往数据中的地学知识。但是其忽视了 GIS 数据特定类的图斑可能包含多种地物类型的情况,即忽视了地理空间和影像空间的不统一[9]。也有研究者从多种分类方法的结合入手,如结合未标定数据的“半监督学习”[10],利用未标定数据的机器学习与标定数据协同分类,提高了样本不

收稿日期: 2012-01-16; **修回日期**:2012-07-20.
基金项目: 国家自然科学基金项目(40971224);国家“863”计划项目(2011AA20101)。
作者简介: 刘锟(1986-),男,邯郸市人,硕士。研究方向为遥感影像智能处理研究。E-mail: liukun@lreis.ac.cn
* **通讯作者**: 杨晓梅(1970-),女,武汉市人,研究员,博士。主要从事遥感与地理信息系统应用研究。
E-mail: yangxm@lreis.ac.cn

足时分类的精度[11]。有研究利用已标定数据来标定未标定数据的聚类中心，结合改进的聚类算法和监督分类器取得了较为理想的分类结果[12]。美国多分辨率土地特征研究项目 MRLC(The Multi-Resolution Land Characteristics)和欧空局发布的 GLC2000(Global Land Cover 2000)项目，也均采用了非监督聚类作为重要分类手段或步骤[2,13-14]。这些研究表明，包含大量前人地学知识的 GIS 数据可以通过样本的选取，间接地作用于影像分类。可见，样本选取能作为二者的结合点。

因此，本文从样本选取着手，以最大似然分类器为例，尝试一种以前期土地覆被为背景数据，结合影像聚类来进行样本自动选取，以改进影像分类的精度。

2 研究数据与方法

2.1 研究区域及数据选取

本文选取的研究区域为广东省珠江口附近，地处东经 113.38°～114.10°，北纬 22.38°～23.05°，研究范围如图 1。研究区内陆部分地势由西北向东南倾斜，地形多样，以丘陵(占 58.68%)为主，兼有平原(占 25.5%)、低山、滩涂等。

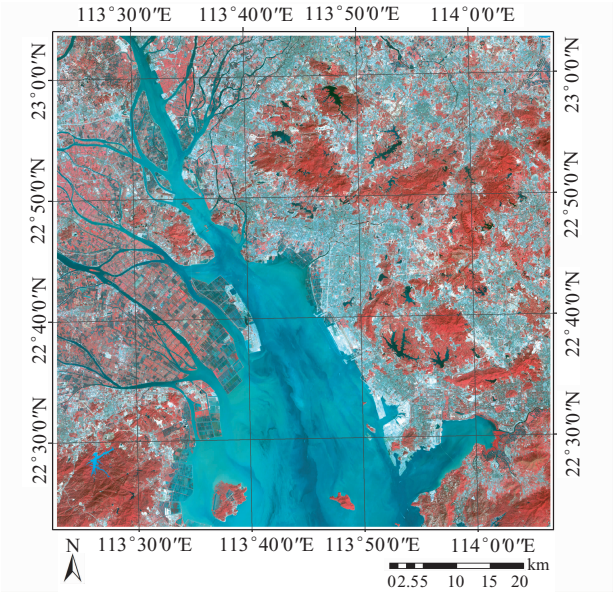

图 1 研究区范围

Fig. 1 The study area

研究数据为 2009 年 11 月 2 日的 TM 影像、该地区 2005 年 DEM 数据(来源于 USGS，http://glovis.usgs.gov/)和 2005 年 1∶5 万海岸带的土地利用解译数据。

2.2 样本选取条件

理论上样本像元应满足 2 个条件：(1)每一类型地物的所有训练样本中像元的实际地物类型应与该地物类型相一致，即所有训练样本中像元的地物类型应为单纯的同一种地物类型，不能是包含不同地物类型的混合像元。(2)选择的样本像元应具有代表性，即训练样本的统计特征量与该类型总体统计特征相接近[5]。为了兼顾样本选取的这 2 个要求，本文采用非监督分类与前期土地覆被数据相叠加的方式，自动获取所需样本，并制定了一个适宜的分类体系，进行样本筛选。提取 TM 影像的前三主成分与坡度信息，作为待分类数据，利用自动选取的样本和人工选取的样本，分别通过最大似然法进行分类，并对 2 种分类结果进行精度对比分析。

3 基于土地覆被数据的训练样本选取

3.1 分类体系转换

在较短时期内，对于同一个研究区，具有一定面积的地物类型一般不会发生根本变化。同时除像元数过少的类型外，特定土地覆被类型对应的影像像元中，与该类型相符的像元占较大比重。对研究区过小(或面积小)、变化过于剧烈的地物类型，本文方法的适用性将受到限制。

我国目前的土地覆被分类体系多是大尺度影像的[15-16]，TM 等中分辨率的土地覆被分类体系较少。TM 影像的空间分辨率为 30m，一般认为其可区分地物类型为一级类[17]。本文综合影像和研究区的特点，参考中国土地资源分类体系和“土地利用现状分类”国家标准[18-19]，拟订了本研究的土地覆被分类体系。

该体系引用了参考系统中水体、林地、灌木林、园地的定义，根据遥感的信息获取特点，参考系统中未利用地细分为裸土地与湿地，城乡工矿居民用地、工矿仓储用地、建设用地等包含城镇生产生活用地的类型，根据其用途和光谱特点分为城镇建设用地和高度开发用地，扩充了参考系统中对于耕地的描述，将城镇绿地也划入其中，类型名称改为耕地及其他人工植被；结合 TM 影像的空间分辨率特

点,扩充混合地表的定义,将在影像上无法识别的植被、土壤和不透水面混合区域划入其中。定义的分类体系如表 1。

将土地利用的 53 个三级类型,根据类别定义和解译标准划入上述 10 类中,生成土地覆被数据。其中难以转化的类别,如其他林地、特殊用地,以及大量观瞻休闲用地等,赋为 0 值,不参与分析。经过对像元的分析,城市绿地、草地等在光谱特征上与耕地差异较小;同时对于研究区,城市所占面积较小,因此,城市绿地的像元数量更为稀少,难以单独区分,本文将城市绿地与耕地合并,作为一个土地覆被类型,命名为耕地及其他人工植被。部分转化关系如表 2。

表 1　研究中的分类体系定义

Tab. 1　Definition of the land cover classes

类型	类别定义
水体	河流、水库坑塘、人工养殖区、浅滩等自然及人工水域
林地	有林地、疏林地等自然及人工林地
高度开发用地	机场、港口码头、工矿等开发强烈、反射率高的区域
城镇建设用地	城镇居民地、城镇道路等城镇建设区域
混合用地	农村、乡镇居民点以及城镇中自然与人工混合区域
园地	各种经济作物人工种植园区
耕地及其他人工植被	耕地、城镇休闲文体等人工植被
裸土地	裸土地、岩砾地等
灌木林	灌丛林地
湿地	沼泽、芦苇地等湿地

表 2　部分土地利用数据与土地覆被数据转化关系

Tab. 2　Conversion of land use types into land cove types

土地利用类型	土地覆被类型	土地利用类型	土地覆被类型
公共建筑用地	高度开发用地	农村宅基地	混合用地
旱地	耕地及其他人工植被	城镇混合住宅	城镇建设用地
裸土地	裸土地	基塘	湿地
其他林地	0	有林地	林地
养殖水面	水体	园地	园地
养殖池塘	水体	灌丛林地	灌木林

人工解译的土地利用类型中,地物斑块包含有多种地物覆被,同时,时相与影像有明显差异。所以,直接将该数据用于样本选取必然会出现大量的影像像元与地物类型不相符的情况,要在其后的样本筛选过程中将其纯化。统计像元数目,其中,灌木林、湿地、裸土地所占比例小于全图的 1%(表 3),难以从影像上分出,与假设前提不符,故剔除上述 3 类。最终确定土地覆被分为水体、林地、高度开发用地、城镇建设用地、混合用地、耕地及其他人工植被、园地 7 类。

表 3　各类型像元所占比例

Tab. 3　The proportion of classes in cells

类型名	像元数	比例(%)
0	47997	3.59
水体	344905	25.83
裸土地	13182	0.99
林地	119292	8.93
灌木林	3394	0.25
湿地	8208	0.61
高度开发用地	233683	17.50
混合用地	43108	3.23
城镇建设用地	134780	10.09
耕地及其他人工植被	31079	2.33
园地	355580	26.63
总计	1335208	100.00

3.2　影像处理与人工样本的选取

本文使用自适应性较好的 ISODATA 非监督聚类方法将影像分为若干聚簇,再以聚簇为单位,利用前期土地覆被数据对聚簇所属类型进行自动标定。聚类结果如图 2。

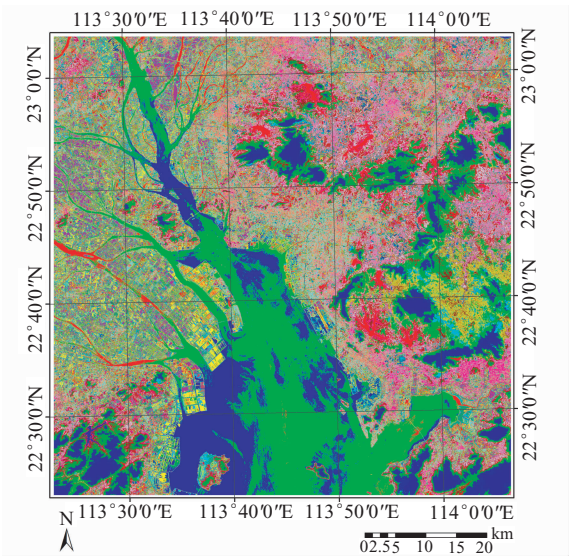

图 2　ISODATA 聚类后结果

Fig. 2　The result of ISODATA

为便于与自动提取样本的分类结果相比较,检验样本的方案是采用分层随机抽样的方式,随机生

成 n 个采样点，辅以多时相的高分辨率影像，对采样点的地物类型作出判断。由 n 个采样点和分类结果建立混淆矩阵，检验分类精度。n 的取值由下式确定[20]：

$$n = \frac{u_1^2 - \alpha/2}{d^2} p(1-p) \quad (1)$$

式中：n 为最少抽样点个数，p 为正确分类的百分比，u 是与 α 对应的置信水平从正态分布概率表上所查的值，d 为误差允许范围。其中，p 可以事先抽取少量样本做精度评估，其结果是取 $\alpha=0.1$ 则 $1-\alpha/2$ 为 95%，对应的 $u=1.96$，本文 d 取 5%。少量抽样表明，p 为 85%左右，则 n 最小值为 195.9，即最少抽取 196 个验证点作为精度验证的样本。

考虑到多波段的影像信息数据量大，波段间相关性较高，同时存在噪声干扰，使用主成分分析，取前三个主成分参与分类[21]。在很多运用 GIS 数据参与分类的研究中，都将坡度参数融入了信息提取中。本实验中，坡度对区分林地和人工植被有明显指示作用，因此，使用 DEM 数据提取坡度数据，与前三个主成分共同参与分类。

3.3 分类样本的自动选取

以行向量 $X=(a, v_1, v_1 \cdots v_n)$ 表示影像上的任一像元，其中，a 为聚类后的类簇号，v_1-v_n 为影像的其他特征，包括坐标，波段 DN 值，DEM 与坡度等。C_k 为前文生成的分类系统中，第 k 类地物所对应的像元集合。统计满足 $X \in C_k$ 的 X 的特征 a 的值，以高度开发用地为例，统计不同的 a 的值所包含的像元数，将 a 的值按照像元数从多到少排列，并设定一个阈值 α，取像元累计百分比不超过 α 的所有 a 为 $A_k=\{a_1, a_1, \cdots, a_n\}$。则第 k 类地物的初步样本为：

$$S_{1k} = \{X \mid\mid X*(1,Z)^T \mid \in A_k\} \quad (2)$$

其中，Z 为行数为 1，列数与像元特征数目 n 相等的零矩阵。

综合研究区土地利用变化强度和实践经验[22]，α 设定为 80%。同样以高度开发用地为例，其聚簇所包含像元累计百分比如表 4。第 25 号聚簇的累计百分比超过 80%，则 $A=\{11,10,27,14,7,16,28,12\}$，即取 11 号到 12 号共 8 个聚簇的像元作为初步样本 S_1。

在此基础上，需要对样本 S_{1k} 再进行一次纯化。设定阈值 β，A_k 中的 a 所对应的像元累计百分比在 β 之内的，视为更能体现类别 k 的特征而与其他类的特征不符的聚簇集合 $B_k=\{b_1, b_2, \cdots, b_m\}$。设除第 k 类之外的类别的 B 的合集为 B_o，则第 k 类地物的最终样本为：

$$S_{2k} = S_{1k} \cap \{X \mid\mid X*(1,Z)^T \mid \notin B_o\} \cup \{X \mid\mid X*(1,Z)^T \mid \in B_k\} \quad (3)$$

其中，Z 为行数为 1，列数与像元特征数目 n 相等的零矩阵。实验中 β 选为 70%。如耕地及其他人工植被和高度开发用地聚类累计像元密度前 80%，如表 5。

表 4 高度开发用地对应聚簇的像元累计百分比

Tab. 4 Cumulative pixel percentage of ISODATA clusters corresponding to locations of highly developed lands

聚簇编号	像元数	累计百分比(%)
11	48939	20.94
10	37929	37.17
27	23647	47.29
14	16898	54.52
7	15960	61.35
16	14443	67.53
28	12429	72.85
12	10686	77.43
25	10409	81.88
……	……	……
18	800	99.06
20	761	99.39
29	641	99.66
19	424	99.84
21	209	99.93
3	56	99.96
22	55	99.98
2	48	100.00

表 5 高度开发用地和耕地及其他人工植被的初步样本

Tab. 5 Initial samples of highly developed land, arable land and other artificial vegetation land

高度开发用地			耕地及其他人工植被		
聚簇编号	像元数	累计百分比(%)	聚簇编号	像元数	累计百分比(%)
11	48939	27.05	14	7275	31.34
10	37929	48.01	16	5589	55.41
27	23647	61.08	9	3839	71.95
14	16898	70.42	11	2497	82.71
7	15960	79.24	8	2345	92.81
16	14443	87.22	25	1669	100.00
28	12429	94.09			
12	10686	100.00			

将其选为样本之后,高度开发用地累计密度不超过 70%的聚类号为 10、11、27,耕地累计密度不超过 70%的聚类号为 14、16、19。高度开发用地中的 11 号聚类在同样也在耕地及其他人工植被的样本中,却不在耕地及其他人工植被的样本累计密度前 70%中,所以将其从耕地及其他人工植被样本中剔除;同理将 14 号聚类从高度开发用地中剔除。而统计耕地及其他人工植被和园地的初步样本,如表 6。其中,累计像元 70%的聚类中,均包含 14 号聚类,则耕地和园地的样本均保留 14 号聚类。

表 6　园地和耕地及其他人工植被的初步样本

Tab. 6　Initial samples of orchard land, arable land and other artificial vegetation land

园地			耕地及其他人工植被		
聚簇编号	像元数	累计密度(%)	聚簇编号	像元数	累计密度(%)
14	65167	24.39	14	7275	31.34
9	57629	45.96	16	5589	55.41
8	50904	65.01	9	3839	71.95
4	34600	77.96	11	2497	82.71
5	31992	89.93	8	2345	92.81
7	26895	100.00	25	1669	100.00

将全部 7 类样本经过上述操作后,选出满足条件的聚簇,将这些聚簇所对应的像元作为最终用于分类的样本。

4　前期土地覆被数据支持下的分类结果与精度分析

将手工选取的样本和自动获取的样本分别导入 ENVI 4.7,对影像前三个主成分和坡度数据使用最大似然分类器进行分类。样本自动选取分类结果如图 3。

分别对 2 个分类结果进行精度评价与比较。利用之前选择的检验样本进行精度评价,其中,自动提取样本的自动分类整体精度为 84.18%,kappa 系数为 0.8066;手工选取样本进行分类的整体精度为 77.04%,kappa 系数为 0.7196。2 种样本获取方法进行分类的混淆矩阵如表 7。

可以看出,样本自动提取方法在整体精度和 kappa 系数上均高于手动选取。具体到各个类别,只有城镇建设用地精度低于手工选取样本的方法。

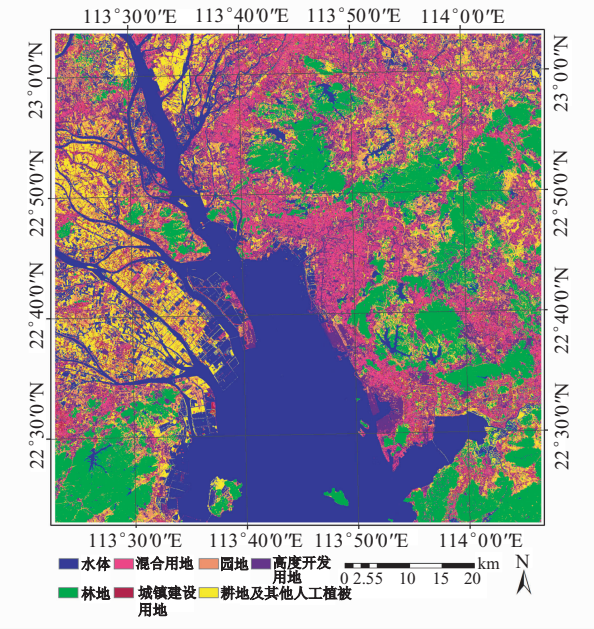

图 3　样本自动选取分类结果

Fig. 3　Result of classification by auto selected samples

从混淆矩阵也可看出,主要错分发生在混合用地、城镇建设用地和高度开发用地 3 类之间,以及园地和耕地及其他人工植被 2 类之间。这主要是由于以下 3 点原因:(1)上述类型之间存在大量区域光

表 7　自动与人工获取样本后分类的混淆矩阵

Tab. 7　The confusion matrix of auto and manual selected samples

(a) 自动获取样本

编号	1	2	3	4	5	6	7
1	93.75	0	0	0	0	0	0
2	0	91.67	0	0	0	0	0
3	0	8.33	60	6.25	12.5	0	0
4	0	0	30	81.25	0	3.85	0
5	1.56	0	10	12.5	68.75	11.54	0
6	0	0	0	0	12.5	69.23	0
7	4.69	0	0	0	6.25	15.38	100

(b) 人工获取样本

编号	1	2	3	4	5	6	7
1	89.06	2.78	0	0	0	0	0
2	3.13	83.33	5	18.75	0	0	0
3	0	5.56	45	0	12.5	0	11.11
4	3.13	2.78	35	68.75	6.25	3.85	0
5	1.56	5.56	15	12.5	56.25	11.54	0
6	0	0	0	0	12.5	73.08	0
7	3.13	0	0	0	12.5	11.54	88.89

注:编号 1 为水体,2 为林地,3 为耕地及其他人工植被,4 为园地,5 为混合用地,6 为城镇建设用地,7 为高度开发用地

谱差异性较弱，在中分辨率尺度上易出现相互影响渗透的现象；(2)高分辨率影像辅助下的检验样本选取，在尺度转换上会出现一定的信息偏差；(3)存在一些地物如果园和部分耕地，难以在影像上准确判别。

5 结论

本文设计了一种依靠前期土地覆被数据，结合ISODATA聚类的样本自动选取方法，通过最大似然分类器将影像分为7个类，分类整体精度达84..18%，kappa系数为0.8066，高于手工选取样本的分类结果。该方法样本的选取是基于待分类影像的，不存在影像的适用性差异。样本的选取过程对分类系统的选择基本没有依赖性，但当分类系统制定不合理时，如类别间可分性较低，则会影响到最终的分类精度。

样本选择过程中，筛选阈值的设置一方面是依赖经验，另一方面是参考当地的土地覆被变化程度。该阈值的敏感度并不强，如本实验中，α的值取在65%～85%之间时，均有相同的分类结果。对于前期数据的选择，最好是相应分辨率的土地覆被或较高分辨率的土地利用数据，这样能最大限度地减少特定地物类型中其他类像元的干扰，同时便于分析处理与精度评价。

本文方法避免了人工选择样本时对混淆区域的误判，是一套依据先验知识、从前期土地调查中获取的样本选取准则。能够获取大量精度较高的可靠样本，节省人力、物力，减少人工干预，从而提升土地利用/土地覆被分类精度。

因该方法依靠于前期数据，故在地物类型变化剧烈，或分布严重不平衡时，该方法的适用性会受到制约。在与特定影像尺度相匹配的土地利用/土地覆被变化研究中，该方法具有自动、高效、可重复等特点，同时分类精度满足要求，可作为一种自动分类方法进一步研究。

参考文献：

[1] 术洪磊，毛赞猷. GIS辅助下的基于知识的遥感影像分类方法研究——以土地覆盖/土地利用类型为例[J]. 测绘学报，1997(4)：328－326.

[2] 张峰，王桥，王文杰，等. 美国高分辨率土地覆盖信息提取技术研究进展[J]. 遥感技术与应用，2008(6)：593－600.

[3] 孙秀邦，范伟，严平，等. 遥感影像土地覆被分类研究进展[J]. 中国农学通报，2007(9)：607－610.

[4] 韩文萍，王金亮，可华明，等. 基于GIS的遥感影像土地利用/土地覆盖信息提取研究——以滇西北香格里拉县为例[J]. 云南地理环境研究，2007(2)：98－102.

[5] 吴健平，杨星卫. 遥感数据监督分类中训练样本的纯化[J]. 国土资源遥感，1996(1)：36－41.

[6] 龚文瑜. GIS辅助遥感影像分类概述[J]. 地理空间信息，2006(2)：15－17.

[7] 肖鹏峰，刘顺喜，冯学智，等. 中分辨率遥感图像土地利用与覆被分类的方法及精度评价[J]. 国土资源遥感，2004(4)：41－45，79.

[8] Borgi A，Akdag H. Knowledge based supervised fuzzy-classification：An application to image processing[J]. Annals of Mathematics and Artificial Intelligence，2001，32(1)：67－86.

[9] 游代安，蒋定华，余旭初. GIS辅助下的Bayes法遥感影像分类[J]. 测绘学院学报，2001(2)：113－117.

[10] Shahshahani B M，Landgrebe D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing，1994，32(5)：1087－1095.

[11] 熊彪，江万寿，李乐林. 基于高斯混合模型的遥感影像半监督分类[J]. 武汉大学学报(信息科学版)，2011(01)：108－112.

[12] Bian X，Zhang T，Fang Z，*et al*. Cluster-based training data preselection and classification for remote sensing images[C]. 2010 IEEE 10th International Conference on Signal Processing (ICSP). Beijing，2010，October.

[13] Vogelmann J，Sohl T，Campbell P and Shaw D. Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources[J]. Environmental Monitoring and Assessment，1998，51(1)：415－428.

[14] Bartholomé E，Belward A. GLC2000：A new approach to global land cover mapping from Earth observation data[J]. International Journal of Remote Sensing，2005，26(9)：1959－1977.

[15] 徐文婷，吴炳方，颜长珍，等. 用SPOT-VGT数据制作中国2000年度土地覆盖数据[J]. 遥感学报，2005，9(2)：204－215.

[16] 刘勇洪，牛铮，徐永明. 基于MODIS数据设计的中国土地覆盖分类系统与应用研究[J]. 农业工程学报，2006，22(5)：99－104.

[17] 张景华,封志明,姜鲁光. 土地利用/土地覆被分类系统研究进展[J]. 资源科学,2011(06):1195 - 1203.
[18] 刘纪远. 中国资源环境遥感宏观调查与动态研究[M]. 北京:中国科学技术出版社,1996.
[19] 陈百明,周小萍. "土地利用现状分类"国家标准的解读[J]. 自然资源学报,2008,22(6):994 - 1003.
[20] Edwards Jr. T C, Moisen G G, Cutler D R. Assessing map accuracy in a remotely sensed, ecoregion-scale cover map[J]. Remote Sensing of Environment,1998,63(1):73 - 83.
[21] 严红萍,俞兵. 主成分分析在遥感图像处理中的应用[J]. 资源环境与工程,2006,20(2):168 - 170.
[22] 高义,苏奋振,孙晓宇,等. 近 20a 广东省海岛海岸带土地利用变化及驱动力分析[J]. 海洋学报,2011,33(4):95 - 103.

Automatic Selection of Classified Samples with the Help of Previous Land Cover Data

LIU Kun[1,2], YANG Xiaomei[1] and ZHANG Tao[1,2]
(1. *Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing* 100101, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: The combination of geographical knowledge and image calibration has long been the principal means of both the traditional visual interpretation and computer automatic classification in remote sensing mapping. Traditional visual interpretation could use the geographic knowledge well because of the artificial participation. However, it goes with the shortcomings that visual interpretation needs a lot of labor and is less efficient. In addition, the computer classification has not applied geographic knowledge in a proper way. Studies have shown that samples as the carrier of geographic knowledge can integrate geographic knowledge into the classification process to some extent. Meanwhile, unsupervised clustering can significantly improve the efficiency of sample selection and solve the problem of scarcity of samples in order to meet the requirement of distribution and purity. These studies provide a basic foundation for integration of geographic knowledge with computer classification. This paper presents an automatic sample selecting method which integrates image clustering with the aid of previous land cover data. The samples were selected automatically based on the TM images by the method mentioned above and used to classify the image later by the maximum likelihood classifier. We also classified the image using the manual samples by the maximum likelihood classifier in order to compare the classified results produced by these two kinds of samples. The test results indicated that the proposed method achieved an overall accuracy of 84.18% and a kappa coefficient of 0.8066 in seven categories, including water body, forest land, orchard and urban construction land. The method proposed in this paper is more efficient than the way of samples selected manually and provides better classification results.

Key words: classification; samples; select automatically; LUCC