维普资讯 http://www.cqvip.com

第36卷 第1期
2007年2月

测 绘 学 报
ACTA GEODAETICA et CARTOGRAPHICA SINICA

Vol.36, No.1
Feb.,2007

文章编号:1001-1595(2007)01-0067-05 中图分类号:P237 文献标识码:A

基于Q型因子分析的训练样本的选择

虞 欣,郑肇葆
(1. 武汉大学 遥感信息工程学院,湖北 武汉 430079)

Selcection of Training Samples Based on Q-Factor Analysis

YU Xin, ZHENG Zhao-bao
(*School of Remote Sensing and Information Engineering, Wuhan University, Wuhan* 430079, *China*)

Abstract: A new method about selction of training samples based on Q-factor analysis is proposed, which can automatically select representative and integral training samples for capturing sufficient samples information and improve classification accuracy. The experimental results demonstrate its feasibility and it performs better than manual selection of training samples.

Key words: factor analysis; training samples; image classification.

摘 要:提出基于Q型因子分析的训练样本的选择方法。它可自动地选择训练样本,而且被选择的训练样本具有代表性,使其在自动分类中充分利用所采集的样本信息,以便得到满意的分类结果。通过实验与分析证明,该方法是可行的,它优于基于人工随机地选择训练样本的方式,而且可获得更高的分类精度。

关键词:因子分析;训练样本;影像分类

1 概 述

航空与遥感影像分类实际上是高度自动化的,但在训练阶段它绝不是自动的,需要人工干预,比如:人为地选择所需的训练样本。监督分类的训练工作不仅是一门科学而且也是一门艺术。它需要图像分析工作者了解研究分类区域内大量的参照数据和完整的地理知识。对于监督分类而言,最重要、最关键的一项是选择训练样本。Hixson认为为了获得很好的监督分类结果,取得有代表性的训练统计量比选择分类的数学模型方法更为重要。而且训练阶段的质量决定着分类阶段的成功与否,因而也体现出从分类结果中所获取信息的价值[1]。可见,训练样本的选择至关重要。

因子分析的对象是变量或样本,如果对象是变量,称为R型分析,也称为主成分分析。如果对象是样本,称为Q型因子分析,也称为主因素分析。主成分分析首先是由Karl Pearson对非随机变量引入的,尔后Hotelling将此方法推广到随机向量的情况。主因素分析则是由Choles Spearman于1900年首先创立的,但两种类型分析并无本质的区别,把研究对象从变量(或特征)转移到样本,便称为Q型因子分析(即把R型因子分析中的数据矩阵X进行转置即可),它是分析存在于多个样本之间关系的一种多元统计方法,其目的在于对一批同属性的样本选出少数的有代表性的样本,来说明由这批样本所携带的样本信息,并由此期望得到科学的单纯性或描述的经济性[2]。

目前在影像分类中,主要是通过人工的方式采集样本,并从中选择所需的训练样本,这种人工的作业方式枯燥、单调,而且选择的结果也因人而异。另外,训练样本的选择(包括比较、修改、合并、剔除等筛选工作)是一项极复杂和费时的工作。为了克服以上的一些不足,进一步提高选择训练样本的自动化程度,本文提出基于Q型因子分析的训练样本的选择方法。它可自动地选择训练样本,且被选的训练样本具有代表性和典型性,其他的样本都可以用这些潜在的"典型公共样本"

收稿日期: 2006-04-11; **修回日期**: 2006-12-13
基金项目: 国家自然科学基金资助项目(40271094;40571102)
作者简介: 虞 欣(1980-),男,浙江义乌人,博士生,主要从事图像解译和人工智能研究。E-mail:china-yuxin@163.com

维普资讯 http://www.cqvip.com

的线性组合表示出来，这样便充分利用所采集的样本信息。作者通过实验与分析证明，基于Q型因子分析的训练样本选择方式是可行的，它优于基于人工随机地选择训练样本的方式，而且可以获得更高的分类精度。

2 基于Q型因子分析的训练样本的选择

2.1 Q型因子模型

因子分析[3]是主成分分析的推广和发展，它也是多元统计分析中降维的一种方法。利用因子分析降维，可以设法找出少数的几个综合因子(少于原来变量或样本个数)来代表众多的因子。而这少数几个综合因子不仅能主要反应原来众多因子的信息，而且彼此独立。Q型因子分析在数学上是基于这样的考虑：假设分别对 n 个样本的 p 个特征(或变量)进行观测，得到一个大小为 $p \times n$ 的原始观测矩阵 $\boldsymbol{X}$。在本文的实验中采用灰度共生矩阵的5个统计性特征和小波分解的8个结构性特征。每个特征就得到 n 个样本的观测值，组成如下的随机观测向量 $\boldsymbol{x}$

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \tag{1}$$

其中，$x_i(i = 1, \cdots, n)$表示某个特征的第 i 个样本的观测值。如果可以用少于 n 个样本(假设 m 个，$m < n$)来代表这组观测样本，那就意味某些样本必然是另一些样本的线性组合。这样就可以减少样本的观测成本，从而简化观测系统。在因子分析模型里，x_i 可用下列的线性关系表示

$$x_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + a_i\varepsilon_i \tag{2}$$

式中，$F_j(j = 1, \cdots, m)$是公因子，它是每个样本中都出现的因子。a_{ij}表示公因子 F_j 的因子载荷，或称之为权系数(即相对重要性)，ε_i 是个别样本所特有的一个特殊因子，a_i 是这一特殊因子的权系数。因子分析模型可简写为如下的矩阵形式

$$\underset{n\times 1}{\boldsymbol{x}} = \underset{n\times m}{\boldsymbol{A}} \times \underset{m\times 1}{\boldsymbol{F}} + \underset{n\times n}{\boldsymbol{a}} \times \underset{n\times 1}{\boldsymbol{\varepsilon}} \text{[4]} \tag{3}$$

式中，$\boldsymbol{A}$ 称为因子 F_j 载荷矩阵，$\boldsymbol{a}$ 为特殊因子载荷，而 $\boldsymbol{F}$ 和 $\boldsymbol{\varepsilon}$ 分别为公因子和特殊因子。

因子载荷矩阵 $\boldsymbol{A}$ 中的元素 a_{ij}表示第 i 个样本与第 j 个公因子的相关系数，根据其绝对值大小就可以判断样本的相对重要性(或典型性)，进而决定是否要把该样本选择为典型样本。$|a_{ij}|$的值越大，就表明第 i 个样本具有较大的载荷，其他的样本无与伦比，因此第 i 个样本就作为公因子 F_j 的代表性样本，这就是训练样本选择的依据。

由于式(3)中除了 $\boldsymbol{x}$ 是已知的，其他的都是未知的，并且上述的分解并不是唯一的，因此Q型因子分析是将 n 个样本之间的相关关系转化为 m 个公因子之间的相关关系，而这个相关关系是通过因子载荷矩阵 $\boldsymbol{A}$ 来反映的。它的初始解可以通过主因子解方法[2](仿R型因子分析的方法)求得 m 个初始因子轴和初始因子载荷矩阵 $\boldsymbol{A}$，但是我们的目的是要知道每个因子轴所代表的意义，它是根据最靠近因子轴的一个特征的样本所决定的。因此，有必要对因子轴进行旋转，使某一因子轴尽可能地靠近一些样本，尽可能地远离另一些样本。这在数学上就是使因子载荷矩阵 $\boldsymbol{A}$ 的结构简化，即使每个因子载荷的平方按列向0,1两极分化，第 j 个主因子的代表性样本在 F_j 因子轴上的载荷系数等于1或趋于1，而在其他因子轴上的系数等于0或趋于0。正交旋转的方法很多，本文采用方差最大正交旋转[3,5]，即使因子载荷矩阵中各因子载荷值的方差 V 达到最大作为因子载荷简化的准则，用公式表示如下

$$V = \sum_{j=1}^{m}\left\{\left[n\sum_{i=1}^{n}\left(\frac{a_{ij}^2}{h_i^2}\right)^2 - \left(\sum_{i=1}^{n}\frac{a_{ij}^2}{h_i^2}\right)^2\right]/n^2\right\} = \max \tag{4}$$

它的优点在于当参加计算的样本数有少量变化时，所得的结果保持不变或变化较小[2]。关于方差最大正交旋转的计算步骤如下

1. 计算公共因子方差。

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 \tag{5}$$

2. 将因子载荷矩阵 $\boldsymbol{A}$ 标准化。

$$b_{ij} = a_{ij}/h_i \tag{6}$$

3. 根据式(4)计算方差 $V_{(k)}$，其中 k 为迭代的次数。

4. 任取因子载荷矩阵 $\boldsymbol{A}$ 中的两列设为 a_{ig}和 $a_{iq}(i = 1, \cdots, n)$，旋转后的矩阵为

$$(a_{ig}, a_{iq})_{(k)} = (a_{ig}, a_{iq})_{(k-1)}\begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix} \tag{7}$$

式中，k 为迭代的次数，j 为样本编号，$g(g = 1, \cdots, m-1)$和 $q(q = g+1, \cdots, m)$为因子编号。

将式(7)代入式(4)，对 φ 求一阶导数，并令其等于零，解出关于 φ 角的方程[2]

$$\tan 4\varphi = \frac{D - 2\times \boldsymbol{A}\times \boldsymbol{B}/n}{C - (\boldsymbol{A}^2 - \boldsymbol{B}^2)\times n} \tag{8}$$

将上述的旋转过程[5]用下式表示为

维普资讯 http://www.cqvip.com

$$\boldsymbol{B} = \boldsymbol{A}\boldsymbol{T}_{gq} \tag{9}$$

式中,$\boldsymbol{T}_{gq}$为相应于式(7)的旋转矩阵。

对因子载荷矩阵 $\boldsymbol{A}$ 中的所有两两成对的因子都进行上述类似的轴旋转计算。如果有 m 个主因子,必须对 $\boldsymbol{A}$ 中所有 m 列全部配对旋转,共有 $C_m^2(=m\times(m-1)/2)$次,全部旋转完毕算一个循环,此时得到的因子载荷矩阵 $\boldsymbol{B}_{(1)}$为

$$\boldsymbol{B}_{(1)} = \boldsymbol{A}\boldsymbol{T}_{12}\cdots\boldsymbol{T}_{1m}\cdots\boldsymbol{T}_{(m-1)m} = \boldsymbol{A}\prod_{g=1}^{m-1}\prod_{q=g+1}^{m}\boldsymbol{T}_{gq} = \boldsymbol{A}\boldsymbol{C}_1 \tag{10}$$

其中,记 $\boldsymbol{C}_1 = \prod_{g=1}^{m-1}\prod_{q=g+1}^{m}\boldsymbol{T}_{gq}$,$\boldsymbol{B}_{(1)}$为对 $\boldsymbol{A}$ 进行正交变换 $\boldsymbol{C}_1$ 而得。经过第一个循环后,可按式(4)计算得 $V_{(1)}$。在第一个循环的基础上从 $\boldsymbol{B}_{(1)}$出发再进行第二个旋转循环,旋转完毕后得 $\boldsymbol{B}_{(2)}$,再按式(4)计算得到 $V_{(2)}$。如此不断重复这个循环,就可得到 V 值的一个非降序列:

$$V_{(1)} \leqslant V_{(2)} \leqslant V_{(3)} \leqslant \cdots \tag{11}$$

由于因子载荷的绝对值不大于 1,故这个序列是有上界的,它必然收敛于某一极限 $\widetilde{V}$。

5. 轴旋转收敛误差的确定。

在旋转过程中,若因子载荷矩阵 $\boldsymbol{A}$ 的总方差 V 值的两次计算之差的绝对值小于某个阈值 ε,则停止旋转。这表明找到了旋转轴的最佳位置,否则继续旋转。这个过程可用下式来表示

$$|V_{(k)} - V_{(k-1)}| < \varepsilon \tag{12}$$

式中一般取 $\varepsilon = 10^{-7}$。

6. 计算旋转后的因子载荷矩阵 $\boldsymbol{G}$。

$$\boldsymbol{G} = (g_{ij}),\quad g_{ij} = b_{ij}\times h_i \tag{13}$$

式中,$i = 1,\cdots,n$,$j = 1,\cdots,m$。

最后,补充说明,在应用因子分析时有一个潜在的要求:在所研究的问题中,要有一定数量的共同因素起作用,即原变量(或样本)之间要具有比较强的相关性。这个条件在选择训练样本中通常是能够满足的。这也可以通过一些统计检验的方法[6]来判断变量(或样本)是否适合作因子分析。

2.2　训练样本的选择方法

Q 型因子分析研究样本之间的相关关系,通过对样本的相关关系矩阵内部结构的研究找出控制所有样本的几个公共因子,因此通过 Q 型因子分析可从一批同属性的样本中挑选出少数并且具有实际意义的、独立的互不相关的、有代表性的典型样本,把选出的这些样本当作这批样本的典型样本,典型样本的信息用于代表(或解释)这批样本所包含的样本信息。当然,其他的未被选出的、剩下的非代表性样本都可以看作是那些典型样本以不同的比例"有机混合"的结果。

本文采用灰度共生矩阵的 5 个统计性特征(角二阶矩、对比度、相关、熵和逆差矩)和小波分解的 8 个结构性特征。对输入图像进行小波(Symlets 小波)分解,分别提取一尺度的近似分量 LL,水平细节分量 LH,垂直细节分量 HL 和对角线细节分量 HH,提取它们各自的均值和方差作为结构性特征。

假设对某类地物采集了 n 个样本,每个样本包括 p 个特征的观测值(在本文的实验中对每个样本都提取上述 13 个同样的特征,即 $p = 13$),得观测矩阵 $\boldsymbol{X}$。则基于 Q 型因子分析的训练样本选择的计算步骤如下:

1. 将观测数据矩阵 $\boldsymbol{X}$ 标准化,得矩阵 $\boldsymbol{S}$[7]。

2. 求样本间的相关矩阵 $\boldsymbol{R}$。

3. 求 $\boldsymbol{R}$ 的特征根和其相应的特征向量。

记 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ 为 $\boldsymbol{R}$ 的特征根,$\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_n$ 为对应的经正交与单位化的标准特征向量。在实际工作中,根据研究对象的情况和特征根累计百分比,取前 $m(m<n)$个主要的特征根。各个特征根的大小代表各个因子的方差在总方差中所占的比例。把特征根所代表的意义分别叫做第一主因素 F_1,第二主因素 F_2,…,第 m 主因素 F_m。主因素累积百分比,又称累积贡献率 η_m。主因素的累积百分比越大,说明选出的主因素拟合原始数据越好,可根据实际研究的问题来确定 η_m 的取值,使得前 m 个特征根之和在 n 个特征根总和中所占的比例达到预先给定的 η_m 取值,如此便可确定 m 的大小,则得到

$$\Lambda = diag(\lambda_1,\lambda_2,\cdots,\lambda_m) \tag{14}$$

$$U^{\mathrm{T}} = (\mu_1,\mu_2,\cdots,\mu_m) \tag{15}$$

$$\mu_i = (\mu_{1i},\mu_{2i},\cdots,\mu_{mi})^{\mathrm{T}} \tag{16}$$

这 m 个公因子可以理解为 m 个独立的"典型公共样本",不同于原始的观测样本。

4. 初始因子载荷矩阵 $\boldsymbol{A}$ 中元素 a_{ij}的计算为

$$a_{ij} = \mu_{ij}\lambda_j^{1/2} \tag{17}$$

从中可看出 μ_i 的分量 μ_{ij}与因子载荷 a_{ij}之间的关系。

5. 按式(10)对初始因子轴进行旋转,直到满足式(12)为止,接着按式(13)计算得到旋转后的因子载荷矩阵 $\boldsymbol{G}$。

维普资讯 http://www.cqvip.com

6. 根据上述的因子载荷矩阵 G,从原始的观测样本中选择出这类地物的代表性样本集(序号)Samples 为

$$\text{Samples}=\{i \mid \max_{1\leqslant i\leqslant n}\{|g_{ij}|\};j=1,\cdots,m\} \quad (18)$$

式中,i 代表原始的观测样本的序号,而 $|g_{ij}|$ 表示对元素 g_{ij} 取绝对值。

7. 重复上述 1~6,选出各类地物的代表性样本集,作为训练样本。

8. 用上述方法选出的训练样本进行训练或学习,得到基于 Naive Bayes Classifiers 的分类器[8,9],接着把所有采集的样本当作测试样本进行测试,最后统计总的分类精度。

3 实验的结果与分析

在本次实验中,从航空影像中提取 141 幅 100×100 像素大小的图像,根据野外调绘的结果,将它们分成 5 种类别属性:灌木、居民地、水田、河流和山地。其中灌木 48 幅(编号 1~48),居民地 33 幅(编号 49~81)、水田 16 幅(编号 82~97)、河流 24 幅(编号 98~121)、山地 20 幅(编号 122~141)。图 1 显示了每类中的一幅样本图像,经过 Q 型因子分析以后,选择的结果用勾(√)的表示被选中的样本,而用叉(×)的表示未被选中的样本。

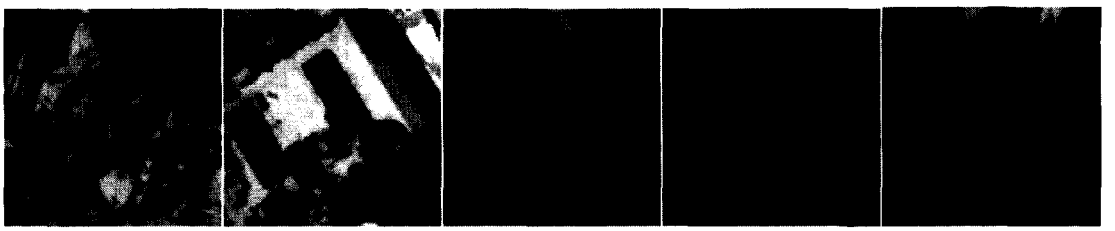

图 1 每一类中的一幅样本图像

Fig. 1 A sample image of each class

从实验中得到如图 2,图 3 和表 1 所示结果。

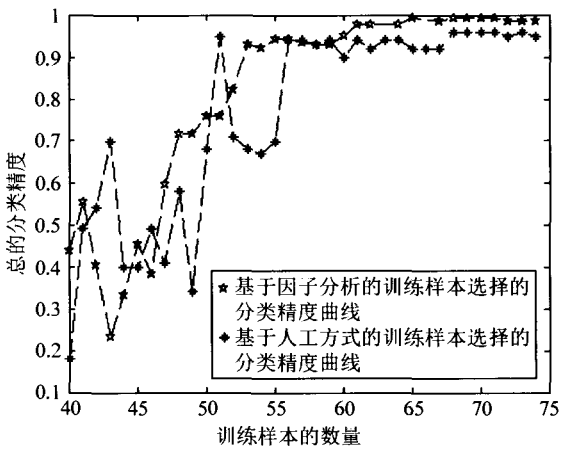

图 2 总的分类精度曲线图

Fig. 2 The Curve of the overall accuracy

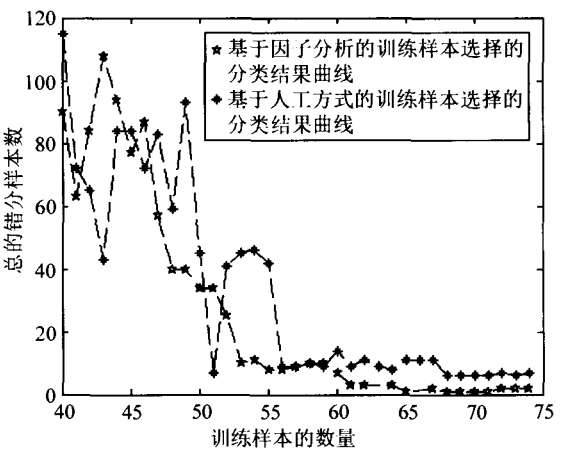

图 3 总的错分样本曲线图

Fig. 3 Tie Curve of the misclassified samples

在图 2,图 3 中横坐标表示训练样本的数量,纵坐标分别表示总的分类精度和错分样本的总数。五角星与米字号分别表示基于因子分析和人工随机方式的训练样本选择后的分类结果。从实验结果中可得到如下结论:

1. 本文提出的方法可以进一步提高分类的精度。从表 1(限于篇幅的原因,表 1 只是列举了累积贡献率在[0.975 0.992]区间内的分类结果)发现当训练样本数多于 53(累积贡献率大于 98%)时,分类的精度可达到 90%以上。在这个基础上再增加训练样本,分类精度的提高比较缓慢。然而,当训练样本数小于 53(累积贡献率小于 98%)时,分类精度明显地偏低,其中最好的也只能达到 82.27%。

维普资讯 http://www.cqvip.com

表 1　分类结果统计表

Tab.1　The statistics of the classification

累计贡献率	训练样本数	总的分类精度	错分样本数
0.975	48	0.716 3	40
0.976	49	0.716 3	40
0.977	50	0.758 9	34
0.978	51	0.758 9	34
0.979	52	0.822 7	25
0.980	53	0.929 1	10
0.981	53	0.929 1	10
0.982	53	0.929 1	10
0.983	53	0.929 1	10
0.984	54	0.922 0	11
0.985	54	0.922 0	11
0.986	55	0.943 3	8
0.987	56	0.943 3	8
0.988	57	0.936 2	9
0.989	58	0.929 1	10
0.990	59	0.929 1	10
0.991	59	0.929 1	10
0.992	60	0.950 4	7

2. 从分类精度来看,本文提出的方法优于基于人工随机的训练样本的选择方法。从图 2,图 3 中,可以看到基于因子分析的训练样本选择的分类精度曲线比基于人工随机的训练样本的选择的分类精度曲线波动较小,而且当训练样本数多于 45 后,前者的曲线呈现递增(图 2)或递减(图 3)的趋势。另外,在图 2 中当训练样本数多于 60 后,前者的分类精度曲线高于后者的分类精度曲线,而且分类精度几乎保持不变。

3. 在实际工作中,需要选择稍多一些的训练样本来弥补主观上的一些不足之处。对于每类中必须具备一定(或者最少)数量的典型(或代表性)样本(临界值),才能充分表达该类样本的总体情况,这些典型样本及它们的线性组合可以不遗漏地充满(或遍历)该类的整个样本空间。当大于临界值时,再增加训练样本,分类精度的提高就显得微乎其微;而当小于临界值时,分类结果就会出现的一些异常或不稳定的现象。

4　结　论

综上所述,基于 Q 型因子分析的训练样本选择方式是可行的,而且它优于基于人工随机的训练样本的选择方式。它使得选出的训练样本具有代表性和典型性,从而可以进一步提高分类的精度。但是,它也有不足之处,因为基于 Q 型因子分析的训练样本的选择,并未同时考虑到变量(或特征)间的相关性,这样人为地将变量和样本割裂开来研究,实际上会损失一些有用的信息。需说明一点,本文只是对所提出的方法进行了初步的尝试,而且本文所得的实验结果和结论只是针对本文的实验条件而言的,对于其他的情况,作者未能进行相关的实验。总而言之,基于 Q 型因子分析的训练样本选择的方法还需要进一步地改进,以便更好地被广泛地应用。

参考文献:

[1] LILLESAND Thomas M. KIEFER Ralph W. Remote Sensing and Image Interpretation [M]. Beijng: Publishing House of Electronics Industry, 2003. (LILLESAND Thomas M, KIEFER Ralph W. 遥感与图像解译[M]. 北京:电子工业出版社,2003.)

[2] Geology Istitute of China Academy of Sciences. Introduction to Geology [M]. Beijing: Geological Publishing House, 1977. (中国科学院地质研究所. 数学地质引论[M]. 北京:地质出版社,1977.)

[3] YU Chong-wen, *et al*. Mathematical Geology and Application [M]. Beijing: Metallurgy Industry Press, 1980. (於崇文,等. 数学地质的方法与应用[M]. 北京:冶金工业出版社,1980.)

[4] SHI Neng. Multivariate Analytical Method Applied to Meteorologically Statistical Prediction [M]. Beijing: Meteorological Press, 1982. (施　能. 气象统计预报中的多元分析方法[M]. 北京:气象出版社, 1992.)

[5] STEEL R G D, TORRIE J H. The Principles and Methods of Statistics[M]. Beijing: Science Press, 1979. (RGD. 斯蒂尔, J H 托里. 数理统计的原理和方法[M]. 北京:科学出版社, 1979.)

[6] ZHANG Yao-ting, FANG Kai-tai. Introduction to Multivariate Statistical Analysis[M]. Beijing: Science Press, (张尧庭, 方开泰. 多元统计分析引论[M]. 北京:科学出版社, 1982.)

[7] WEI Ke-rang, JIANG Cong-shi. Error Analysis of Spatial Data[M]. Beijing: Science Press, 2003. (魏克让, 江聪世. 空间数据的误差处理[M]. 北京:科学出版社, 2003.)

[8] YU Xin, ZHENG Zhao-bao, TANG Lin, YE Zhi-wei. Aerial Image Texture Classification Based on Naive Bayes Classifiers [J]. Geomatics and Information Science of Wuhan University, 2006, 31(2): 108-111. (虞欣, 郑肇葆, 汤凌, 叶志伟. 基于 Naive Bayes Classifiers 的航空影像纹理分类[J]. 武汉大学学报(信息版), 2006, 31(2): 108-111.)

[9] YU Xin, ZHENG Zhao-bao, LI Lin-yi, YE Zhi-wei. Texture Classification of Aerial Image Based on PCA-NBC[A]. The Fourth International Symposium on MIPPR[C]. Wuhan: [s. n.], 2005.

[10] ZHENG Zhao-bao. Classification of Image Texture by Combination of Genetical Algorithm and Simplex Method[J]. Acta Geodaetica et Cartographica Sinica, 2003, 32(4): 325-329. (郑肇葆. 遗传算法与单纯形法组合的影像纹理分类方法[J]. 测绘学报, 2003, 32(4): 325-329.)

[11] NIXON M, AGUADO A. Feature Extraction and Image Processing [M]. [s. l.]: Butterworth-Heineman Linacre House, 2002.

(责任编辑:丛树平)