

遥感图像土地覆盖分类中多源特征数据选择研究

张元, 陈亮, 王文种, 王军战

(河海大学水文水资源及水利工程国家重点实验室, 南京 210098; 河海大学土木工程学院, 南京 210098)

【摘 要】多源特征数据可以提高遥感图像的分类精度, 选择合适的特征数据十分重要。利用基尼指数对多尺度纹理信息、主成分变换前三分量、地形数据等特征进行选择, 选出最佳特征子集。利用支持向量机、神经网络分类法、最大似然法分别对全部特征数据和最佳特征子集结合多光谱数据进行分类。实验结果表明: 基尼指数可以有效地对多源特征数据进行选择, 特征选择可以提高分类器效率, 提高分类精度。

【关键词】基尼指数; 多源特征数据; 特征选择; 分类

【中图分类号】TP751

【文献标识码】A

【文章编号】1009-2307(2009)02-0055-03

DOI: 10.3771/j.issn.1009-2307.2009.02.018

1 引言

利用遥感手段获取土地利用/土地覆盖信息是遥感应用的一个重要领域, 其中分类是重要环节之一。利用多特征和辅助数据提高分类精度越来越被国内外学者所重视, 张锦水^[1]利用纹理和结构等信息对 IKONOS 高分辨率图像进行分类, 杜明义^[2]、William^[3]将光谱信息结合植被指数、高程等信息进行土地分类, 宋翠玉^[4]运用光谱信息和多尺度纹理进行城市扩张变化监测, 彭光雄^[5]利用纹理分析方法对 CBERS02 卫星图像进行土地覆盖信息提取, Coburn^[6]利用多尺度纹理提高林地分类精度, 邓小炼^[7]结合主成分变换、缨帽变换、植被指数等特征进行土地利用分类。这些研究表明, 利用光谱信息结合纹理、植被指数、高程等信息进行分类能够增加地物的可分性, 提高分类精度。但是, 诸多研究关注的是多源特征对分类精度的影响, 而对多源特征数据的选择、以及特征选择对分类精度影响的研究尚不多见。随着参与分类数据的维数增加, 一方面计算量增加, 分类时间会增加; 另一方面, 多源特征数据中的噪声也可能影响分类精度。因此, 针对多源特征数据进行选择, 提取有效的特征参与分类有着重要意义。

特征选择方法大致可以分为两种: 过滤方法(Filter)和打包方法(Wrapper)。过滤特征选择方法是一种计算效率较高的方法。目前使用最多的过滤特征选择方法有概率距离和相关测量法、类间和类内距离测量法、信息熵法等^[8]。然而, 多源特征数据的相关性相对较小, 因此我们只需更多考虑类间可分性。但是当数据维数较大时, 可能的波段组合数较多, 类间和类内距离测量法计算量将很大。基尼指数(Gini Index)是一种常用的决策树分裂算法, 已成功应用于 CART 算法、SLQ 算法、SPRINT 算法等。王圆圆^[8]研究表明决策树不但是一种分类算法, 还可以用来特征选择。Charu^[9]、尚文倩^[10]等采用基尼指数来解决文本分类中的特征选择问题, 并取得了较好的效果。因此, 本文将基尼指数引入遥感图像分类中的多源特征数据选择, 并利用支持向量机、神经网络分类和最大似然法进行分类,

从而分析比较本文提出的多源特征数据选择方法的有效性。

2 基尼指数

基尼指数被作为一种不纯度指标用于决策树模型中来选择最佳测试变量和分割阈值, 基尼指数的数学定义如下^[10]:

$$Gini(t) = 1 - \sum_{j=1}^J p^2(j|h)$$

$$\text{其中, } p(j|h) = \frac{n_j(h)}{n(h)}, \sum_{j=1}^J p(j|h) = 1。$$

式中, $p(j|h)$ 是从训练样本集中随机抽取一个样本, 当某一特征的特征值为 h 时属于第 j 类的概率, $n_j(h)$ 为训练样本中该特征特征值为 h 时属于第 j 类的样本个数, $n(h)$ 为训练样本中该特征的特征值为 h 的样本个数, J 为类别个数。

如果根据某个特征进行分割, 将训练样本分为 k 个子集, 那么进行这个分割的 Gini 为:

$$Gini(T) = \sum_{i=1}^k \frac{n_i}{n} Gini(i)$$

k 是子集的个数, n_i 是子集 i 的样本数, n 是样本总数。

基尼指数特征选择的基本思想就是利用训练样本对全部特征进行遍历, 计算每个特征的基尼指数值, 若能提供最小的基尼指数值, 就被选作为最佳特征。对于某一特征, 如果各类别样本在该特征空间分布均匀, 则具有较大的基尼指数值, 不利于样本分割; 如果各类别样本在特征空间聚集在某几个区域, 则具有较小的基尼指数值, 有利于样本分割。因此, 可以用基尼指数来选择一组对样本具有最好区分能力的特征。

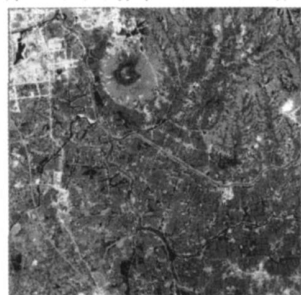


图 1 研究区原始影像
(RGB 波段组合)

3 研究区概况及研究数据

研究区位于南京市郊, 近年来城市扩张迅速, 土地利用类型变化较大。研究区内各类地物类型交错分布, 地块较为破碎。研究区的主要地物类型有: 水田、旱地、林地、草地、水体、居民及交通过地。水田主要分布在河谷, 黄土丘陵以及缓岗以旱作为主, 丘陵山地中林地和草地交错分布。本次研究采用的研究数据为 2002 年 ASTER 数据的 14 个波段及 DEM 数据。

4 特征选择

4.1 特征构建

遥感图像不但包含光谱信息, 还包含有空间结构信息。



作者简介: 张元(1982-), 男, 硕士研究生, 贵州省都匀市人, 从事 GIS 和遥感研究。

E-mail: zhangyuan_zsu@163.com

收稿日期: 2007-11-09

纹理是一种常用的空间结构信息，它不仅反映了图像的灰度统计信息，而且反映了地物本身的结构特征和地物空间排列的关系，是进行目视解译的重要标志之一^[11]。此外，地物具有多尺度性，很难用一个尺度描述一幅影像的所有地物纹理特征^[12]，因此多尺度纹理能够更好地描述地物。灰度共生矩阵法是目前公认的一种比较成熟、有效的提取纹理的方法，它是 Haralick提出的一种纹理描述方法^[13]。本文对 ASTER 多光谱数据进行主成分变换，选取第一主成分进行纹理信息提取。采用 8 个基于灰度共生矩阵的描述纹理特征的统计量来计算纹理，它们分别为均值、方差、均匀性、对比度、非相似度、熵、角二阶矩、相关度。纹理移动方向取 0°、45°、90°、135 这 4 个方向的平均值，纹理的步长取 1，窗口大小为 3×3、5×5、7×7、9×9、11×11。

地物在空间上的分布受地域自然条件的控制和人为因素的干预，往往存在某种地域分异规律。因此，本文将 DEM，以及由 DEM 生成的坡度、坡向图作为地形特征加入特征选择。此外，主成分变换的前三分量及归一化植被指数也经常被作为特征进行信息提取，本文也将其加入特征集。将 ASTER 多光谱及全部特征归一化到 0-1 之间进行特征选择和分类。

4.2 特征选择

本文将该地区 SPOT 卫星的全色波段和多光谱波段的融合数据作为参考数据，在 Aster 图像上选取了 3100 个样本点，其中，2100 个为训练样本，1000 个为验证样本。根据训练样本计算多源特征数据的 Gini 指数值，见表 1。

表 1 各特征基尼指数值

特征	Gini值	特征	Gini值	特征	Gini值	特征	Gini值	特征	Gini值
M3	0.37	H5	0.65	D7	0.66	A9	0.66	PC1	0.37
V3	0.76	Con5	0.71	E7	0.64	Cor9	0.75	PC2	0.48
H3	0.67	D5	0.66	A7	0.67	M11	0.47	PC3	0.66
Con3	0.75	E5	0.66	Cor7	0.76	V11	0.67	NDVI	0.44
D3	0.71	A5	0.67	M9	0.46	H11	0.65	DEM	0.52
E3	0.72	Cor5	0.78	V9	0.69	Con11	0.67	Slope	0.67
A3	0.73	M7	0.43	H9	0.64	D11	0.67	Aspect	0.49
Cor3	0.77	V7	0.69	Con9	0.68	E11	0.63		
M5	0.4	H7	0.64	D9	0.67	A11	0.65		
V5	0.72	Con7	0.69	E9	0.64	Cor11	0.74		

其中 M、V、H、Con、D、E、A、Cor 分别为均值、方差、均匀度、对比度、非相似度、熵、角二阶矩、相关度，字母后面的数字代表纹理窗口大小。

从表 1 可以看出，对于纹理而言，均值、均匀度以及熵纹理量的基尼指数值都比较小，表明各地物在这几个纹理上的可分性较好。进一步分析可知，水田、旱地以及草地在光谱特征空间上有较大混淆，但在纹理上有较好的可分性。但是水田斑块较大，均匀度大，熵值小；旱地比较破碎，均匀度较大，熵值大；与水田和旱地不一样，草地为天然植被，其覆盖度差异较大，均匀度较小，熵值大。所以，均匀度和熵对这三种地物有很好的区分能力。而均值纹理量与原始数据有很大的相关性，对地物也具有较好的区分能力。因此，均值、均匀度以及熵的基尼指数值较小。由此可知，基尼指数的选择结果与实际情况基本相符，说明基尼指数进行多源特征数据选择是可行的。表 1 中，主成分第一、二分量，植被指数，DEM 以及坡向都具有较小的基尼指数值。鉴于上面分析结果，本文选择基尼指数值小于等于 0.65 的特征进行分类。

4.3 分类与分析

为了进一步说明本多源特征数据选择对分类精度的影响，选用支持向量机（SVM）、神经网络分类法（bp-NN）、最大似然法（MLC）对全部特征（ASTER 多光谱及全部特征）

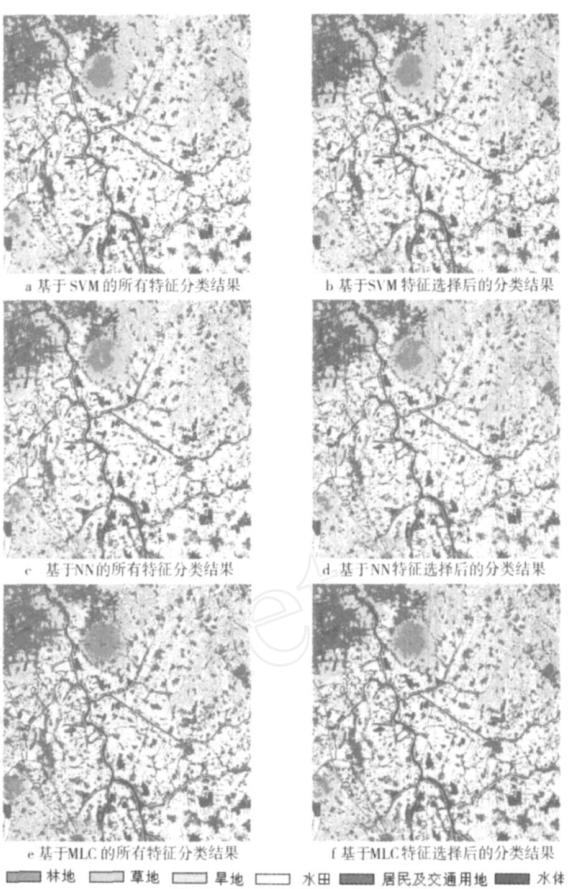


图 2 各分类器特征选择前后的分类结果

及最佳特征子集（ASTER 多光谱及最佳特征子集）进行分类。支持向量机选用的是多类 SVM 分类器，核函数选择径向基函数，参数确定采用的是交叉验证法，用于全部特征和最佳特征子集分类的支持向量机惩罚因子 C 分别为 75 和 120，核函数宽度 分别为 0.3 和 0.16。神经网络选用的是带有两个隐含层的 BP 神经网络，通过反复训练得到其参数，学习率和动量因子分别为 0.1 和 0.5，循环数为 5000。各分类器分类结果见图 2，分类精度见表 2。本文还对各分类器的运行时间做了统计比较，见表 2。

表 2 特征选择前后各分类器效率比较

分类器	全部特征			特征选择		
	总体精度	Kappa系数	运行时间	总体精度	Kappa系数	运行时间
SVM	88.5	0.83	8 24	90	0.87	3 48
NN	83.2	0.77	63 4	86.4	0.81	25 36
MLC	77.2	0.69	4 16	82	0.76	1 37

从表 2 结合图 2 可以看出，经特征选择后各分类器运行的时间减少一半之多（而基于 Gini 指数特征选择的时间仅为 1），而且各分类器的精度都有所提高，支持向量机、神经网络和最大似然分类法分别提高了 1.5%、3.2% 和 4.8%。进一步研究发现，特征选择后各分类器分类结果的图斑块大小更加合理，地物的细节信息提取得更加准确，这是因为通过特征选择，较好地去除了一些对分类贡献很小的噪声信息，减少了噪声对分类结果的影响。由此可见，对多源特征数据进行选择用于进一步分类，在提高分类效率的同时还能提高分类精度。这表明对多源特征数据进行特征选择是有必要的，实验结果还表明基尼指数能够有效地进行特征选择。

从表 2 还可以发现，支持向量机分类精度最高，而且其受特征选择的影响较神经网络和最大似然分类的要小，

这是因为支持向量机有较强抗干扰能力和很好的泛化能力。神经网络的分类精度较高,但是其训练的时间最长,远远多于其他两个分类器,这与其分类机制有关。由此可以看出,支持向量机是一种理想的分类器。

5 结束语

本文利用基尼指数对由 ASTER 数据派生的纹理等特征及其他地理辅助数据进行特征选择,采用支持向量机、神经网络和最大似然分类法进行分类比较,得出以下结论:基尼指数能够有效应用于多源特征选择,选择结果与实际情况相符合,基于基尼指数的特征选择方法计算量相对较小,是一种方便可行的特征选择方法;对多源数据特征进行选择,选择合适的特征进行分类,能够有效地提高分类效率,提高分类精度;支持向量机的分类精度最高,抗干扰能力强,同时分类时间也较短,是一种有效的分类器。

基尼指数虽然对多源特征数据选择是可行的,但是对于波段信息冗余较大的高光谱数据而言还有待进一步验证。

参考文献

- [1] 张锦水,何春阳,潘耀忠等.基于 SVM 的多源信息复合的高空间分辨率遥感数据分类研究[J]. 遥感学报,2006, 10(1): 49-57.
- [2] 杜明义,武文波,郭达志.多源地学信息在土地荒漠化遥感分类中的应用研究[J]. 中国图象图形学报,2002, 7(7): 740-743.
- [3] William L Stefanov, Michael S, Ramsey, Phil PR, Christensen. Monitoring urban land cover change: An expert system approach to land cover classification of semi-arid to arid urban centers [J]. Remote Sensing of Environment, 2001, 77(2): 173-185.
- [4] 宋翠玉,李培军,杨锋杰.运用多尺度图像纹理进行城市扩展变化检测 [J]. 国土资源遥感,2006, (3): 37-42.
- [5] 彭光雄,李京,何宇华,等.利用纹理分析方法提取 CBERS02 星 CCD 图像土地覆盖信息 [J]. 遥感技术与应用,2007, 22(1): 8-13.
- [6] Coburn C A, Roberts A C B. A multiscale texture analysis procedure for improved forest stand classification [J]. International Journal of Remote Sensing, 2004, 25 (20): 4287-4308.
- [7] 邓小炼.基于变化矢量分析的土地利用变化监测方法研究 [D]. 北京:中国科学院遥感研究所, 2006.
- [8] 毛勇.基于支持向量机的特征选择方法的研究与应用 [D]. 浙江:浙江大学, 2006.
- [9] 王圆圆,李京.基于决策树的高光谱数据特征选择及其对分类结果的影响分析 [J]. 遥感学报,2007, 11(1): 69-76.
- [10] C Charu, et al. On the merits of building categorization system by supervised clustering [C]. The 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Cal, 1999: 352-356.
- [11] 尚文倩,黄厚宽,刘玉玲.文本分类中基于基尼指数的特征选择算法研究 [J]. 计算机研究与发展,2006, 43(10): 1688-1694.
- [12] 陈春,等.遥感信源色彩信号的提取与复现 [J]. 测绘科学,2006, 31(1) .
- [13] ATKINSON, P.M., AHLN, P.. Spatial variation in land cover and choice of spatial resolution for remote sensing [J]. International Journal of Remote Sensing, 2004, 25(18): 1351-1364.
- [14] Haralick R M. Statistical and structural approaches to texture [J]. IEEE Proceedings, 1979, 67(5): 786-804.

Multi-source feature data selection for land cover classification using remote sensing image

Abstract: Multi-source feature data can be used to improve the accuracy of land cover classification. However, selecting suitable features is an important step. Gini index was applied to select features from the feature set including multi-scale texture, the components of principal component analysis, and terrain data. The selected features and multi-band data were classified into six classes by support vector machine, neural network classifier and maximum likelihood classifier. The results showed that Gini index could select features successfully, and the classification accuracies were improved while the run time was reduced after feature selection.

Key words: Gini index; multi-source feature data; feature selection; land cover classification

ZHANG Yuan, CHEN Liang, WANG Wen-zhong, WANG Jun-zhan (State Key Laboratory Of Hydrology-Water Resource And Hydraulic Engineering of Hohai University, Nanjing 210098, China; College of Civil Engineering, Hohai University, Nanjing 210098, China)

(上接第 27 页)

- [4] 余璟明,何希琼,程冬爱.基于离散小波变换的时间序列数据挖掘 [J]. 计算机应用,2005, 25(3) .
- [5] 连达军,袁铭.基于极大熵谱估计准则的动态数据预测方法及应用 [J]. 苏州科技学院学报,2005, 3.
- [6] 徐昕,李涛,伯晓晨.Matlab 工具箱应用指南—控制工程篇 [M]. 北京:电子工业出版社, 2000: 32-167.
- [7] 李毅敏,尹秉坤.基于极大熵谱估计的短时间序列分析与预测方法的研究 [J]. 武汉科技大学学报, 2005, 28(30): 264-265.
- [8] 郭英,等.基于时间序列趋势外推改进模型的 SISE 预测 [J]. 测绘科学,2007, 32(1) .

Research on time series prediction based on maximum entropy and wavelet

Abstract: Based on maximum entropy principle, the recorded deformation data of Tiger Hill pagoda of Suzhou is taken as an example to predict time series data. This paper analyzes the influence of different strategies of data sampling on the precision of prediction. According to the same sampling data, this paper concludes different prediction curves based on different methods of parameter estimation for various models, and compares these curves through de-noised disposal by wavelet analysis. The result indicates that the periods of deformation for pagoda between 1985 and 2000 have extended gradually, and the whole situation of deformation has being intensified. The prediction curve of AR(p) model based on maximum entropy principle can fit real curve very well, symmetrically sampled data is key to enhance the precision of model prediction.

Key words: time series; maximum entropy; AR(p) model; wavelet

JING Kang, ZHANG Yong-zhan, LIAN Da-jun, MIN Feng-yang (Key Laboratory of Coast and Island Development of MOE, Nanjing University, Nanjing 210093, China; School of Environment Science and Engineering, Suzhou University of Science and Technology, Suzhou 215000, China)