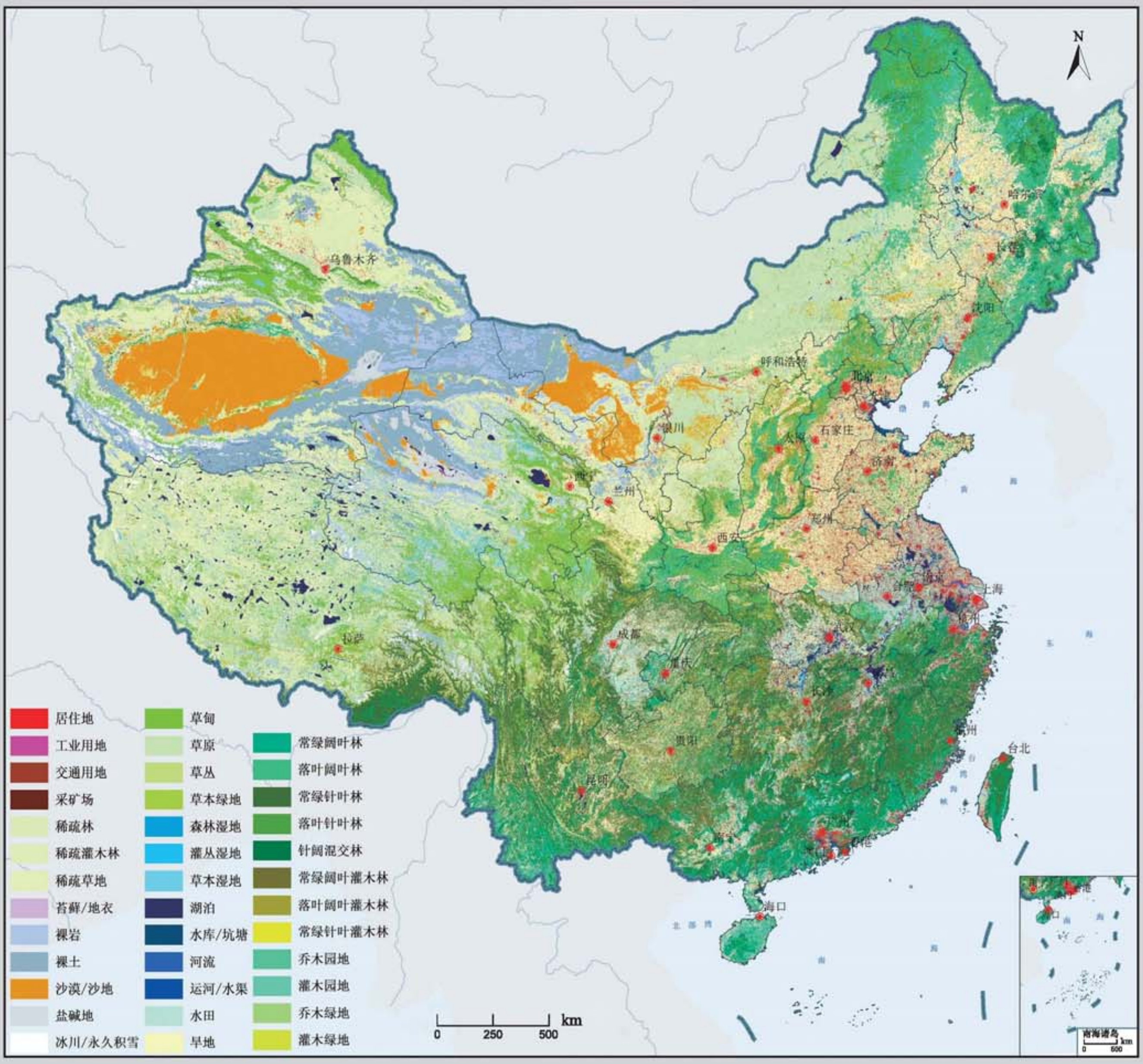
科学出版社
中国地理学会环境遥感分会 出版
中国科学院遥感与数字地球研究所 主办

JOURNAL OF REMOTE SENSING 遥感学报

2013年 Vol.17 第17卷 No.4 第4期 ISSN 1007-4619 CN11-3841 / TP CODEN YXAUAB

2010年中国土地覆被遥感监测数据集（ChinaCover2010）



百种中国杰出学术期刊

遥 感 学 报

Yaogan Xuebao

第 17 卷　　第 4 期　　2013 年 7 月

目　　次

JOURNAL OF REMOTE SENSING

(Vol. 17 No. 4 July, 2013)

CONTENTS

(to be continued to Inside Back Cover)

(continued from Contents page)

Haze: Remote Sensing

Feature selection and its application in object-oriented classification

WANG He, CHEN Jinsong, YU Xiaomin

Shenzhen Institutes of Advanced Technology, Chinese Academy of Science,
Shenzhen 518055, *China*

Abstract: The SEaTH method (SEparability and THresholds) can select features and compute thresholds automatically based on Jeffries-Matusita distance to separate classes, which may cause information redundancy and affect classification results. In this paper, a new method based on SEaTH is proposed, and then the method that considers the Jeffries-Matusita distance, the inner class distance, and the correlation of different features is used in the classification of multi-resolution remote sensing images of Zhaoqing, China. The result indicates that our method can select more effective information than the original SEaTH algorithm, with accuracy in extracting farmland improved by 12.26%, and total accuracy improved from 80% to 85.26%. Such an improvement is significant to the classification when multi-temporal data are difficult to be obtained.

Key words: object-oriented classification, feature selection, SEaTH algorithm, land cover classification, feature decorrelation

CLC number: P237 **Document code:** A

Citation format: **Wang H, Chen J S and Yu X M. 2013. Feature selection and its application in object-oriented classification. Journal of Remote Sensing, 17(4): 816–829 [DOI: 10.11834/jrs.20132257]**

1 INTRODUCTION

Land-cover classification is a major application of remote sensing, and is also the foundation of studying climate, ecology, and the environment on a regional scale. With the development of the social economy and the progress of science and technology, problems such as global climate change and ecological environment destruction have become more prominent, and have brought serious consequences to human survival and progress. Obtaining land cover information fast, accurately, and timely to provide reliable basic data for ecological environment protection and economic development planning is an urgent problem. Simultaneously, selecting the effective classification features to improve the classification accuracy and increase separable feature types is also a key concern. Along with in-depth research, a greater demand exists for the method and key technology of land-cover classification to obtain more accurate and detailed surface information in an efficient manner.

Today, the features of optical remote sensing images used for classification are varied, and the most basic ones are spectral features, texture features, geometric characteristics, and so on. With the increase of remote sensing data sources and the method of classification, the basic features of multi-source remote sensing data have further evolved, thus deducing more information that is available. More information can be used in object-oriented classification than in pixel-based classification, particularly shape and context information, which can explore features from different aspects to improve separability and accuracy (Laliberte, et al., 2012). However, given the increasing types and dimensions of the features, it is difficult to obtain accurate features and their reasonable thresholds quickly. To solve these problems, several parameter selection methods have been proposed (Koller, et al., 1996; Feng, et al., 2008; Gao, 2010; Bruzzone, et al., 2000; Zhang, 2006; Ye, 2007; He & Xu, 2002; Peng, et al., 2005; Nussbaum, et al., 2006; Zheng, 2010; Addink, et al., 2010; Chen, et al., 2010; Han, et al., 2010; Wu, et al., 2009). Among these methods, SEaTH is a typical algorithm used in feature selection and threshold computation based on the Jeffries-Matusita distance and the Gaussian probability mixture model, as proposed by Nussbaum, et al. (2006). Previous studies have shown that this algorithm with higher efficiency and less manual intervention provides better results than the nearest neighbor classification and the maximum likelihood classification (Gao, et al., 2011). The SEaTH algorithm considers only the probability distance but ignores the correlation between the features and the inner differences of the classes, which can cause information redundancy and affect the classification results.

In this paper, a new method considering the correlation, inner distance, and probability distance is proposed to select the optimal classification features. Then, Landsat TM and HJ-1A/1B images of Zhaoqing city are used in land-cover classification research by object-oriented classification.

Received: 2012-09-11; **Accepted:** 2013-01-08; **Version of record first published:** 2013-01-15

Foundation: Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA05050107-03)

First author biography: WANG He (1986—), female, master. Her current research interests are information extraction from multi-source remote sensing images and its applications. E-mail: he.wang@siat.ac.cn

Corresponding author biography: YU Xiaomin (1982—), female, Ph.D., assistant researcher. She specializes in dynamic change detection of remote sensing. E-mail: yu.x.min@gmail.com

2 METHODOLOGIES

2.1 SEaTH algorithm

The feature-analyzing tool SEaTH identifies the characteristic features with a statistical approach based on training objects. Its main steps are listed below.

(1) Construction of the training dataset. The training dataset, which contains all types of features, is built (the sample number accounts for approximately 2% of all image objects). Then, the candidate features are presented.

(2) Selection of the optimum classification features. For a given training dataset, the probability distribution for each class can be estimated and used to calculate the Jeffries-Matusita distance between two object classes, C_1 and C_2, as their separability. Under the assumption of normal probability distribution, the Jeffries-Matusita distance (J) is given by

$$J = 2(1 - e^{-B}) \tag{1}$$

$$B = \frac{1}{8}(m_1 - m_2)^2 \frac{2}{\sigma_1^2 + \sigma_2^2} + \frac{1}{2}\ln\left(\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}\right) \tag{2}$$

where m_i and σ_i^2, $i=1,2$, are the mean and the variance, respectively, for the two feature distributions. The scale of J is $[0—2]$ in terms of B. The higher J is, the better the separability is and the less the misclassified objects are. Thus, the features with higher J were selected for classification.

(3) Calculation of the threshold. When a Gaussian probability mixture model is fit to the frequency distribution of a feature for two object classes C_1 and C_2, the decision threshold that minimizes the error probability can be obtained by

$$p(x \mid C_1)P(C_1) = p(x \mid C_2)p(C_2) \tag{3}$$

where $p(x|C_1)$ is a normal distribution with mean m_{c1} and variance σ_{c1}^2 and similarly for $p(x|C_2)$. Then, the threshold can be calculated by

$$x_{1(2)} = \frac{1}{\sigma_{c1}^2 - \sigma_{c2}^2}\Big[m_{c2}\sigma_{c1}^2 - m_{c1}\sigma_{c2}^2 \pm \sigma_{c1}\sigma_{c2}\sqrt{(m_{c1} - m_{c2})^2 + 2A(\sigma_{c1}^2 - \sigma_{c2}^2)}\Big] \tag{4}$$

$$A = \log\left[\frac{\sigma_{c1}}{\sigma_{c2}} \times \frac{p(C_2)}{p(C_1)}\right] \tag{5}$$

The one between m_{c1} and m_{c2} is used as the final threshold.

2.2 Improved feature selection method

A significant correlation between classification features is observed, which can cause information redundancy. Based on the theory that adding or deleting related features from the original feature set does not affect classification ability (Ji, et al., 2001), the decorrelation is introduced in the method discussed in this paper. The other improvement is that the combination of inner distance and the probability distance is used to build the feature selection index instead of the probability distance only. The improved procedure is:

(1) Feature decorrelation. If the original multidimensional feature set is $F_N=(f_1, f_2, \cdots, f_N)$ and the number of the objects is K, their correlation coefficient matrix is given by

$$\boldsymbol{r} = \begin{pmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & & \vdots \\ r_{N1} & \cdots & r_{NN} \end{pmatrix} \tag{6}$$

$$r_{ij} = \frac{\sum_{l=1}^{K}(x_i^l - \bar{x}_i)(x_j^l - \bar{x}_j)}{\sqrt{\sum_{l=1}^{K}(x_i^l - \bar{x}_i)^2 \sum_{l=1}^{K}(x_j^l - \bar{x}_j)^2}} \quad (i = 1, \cdots, N; j = 1, \cdots, N) \tag{7}$$

where x_i^l is the value of the i^{th} feature of the l^{th} object and $\bar{x}_i$ is the mean of the ith feature that can be identified by $\bar{x}_i = \frac{1}{K}\sum_{l=1}^{K} x_i^l$.

According to the criterion of image information measurement that the smaller the variance is, the less the information included in the feature is, and the larger the variance is, the more the information is (Zhang, et al., 2009), the features with higher correlation and smaller variance are removed. Under different threshold conditions, the number of the selected features is different. Thus, the smaller the threshold is, the less the selected features are. In this paper, the threshold is cited at 0.95 by experience. The decorrelation feature set $\boldsymbol{F}_M$ ($M<N$) that contains most information of the original features and avoids information redundancy is the basis for the next step.

(2) Probability distance calculation. The probability distance indicates the distance between two classes and can be obtained by Eq.(1). Ten better features for distinguishing each two classes are selected by descending their J values and the first ten are selected to compose the feature set $\boldsymbol{F}_p$ as the basis of the next step.

(3) Feature normalization. The orders of magnitudes of the selected features are different, so in order to be compared and calculated, the features should be normalized to 0 to 1 before the inner distance is calculated. The formula is given by

$$F' = \frac{F - F_{\min}}{F_{\max} - F_{\min}} \tag{8}$$

(4) Inner distance calculation. For the normalized feature set $\boldsymbol{F}_p' = (f_1, f_2, \cdots, f_p)$, the classes set is $\boldsymbol{C}_n = (c_1, c_2, \cdots, c_n)$, and the number of the typical objects is $\boldsymbol{K}_n = (k_1, k_2, \cdots, k_n)$, so the inner distance can be obtained by

$$d = \frac{2}{k(k-1)}\sum_{l=1}^{k}\sum_{m=l+1}^{k}(x_l - x_m)^2 \tag{9}$$

where x_l and x_m are the values of the l^{th} and the m^{th} sample features of a given class, and k is the number of the samples. The inner distances d_1 and d_2 are calculated, and according to the numbers of class C_1 and C_2, the weighted inner distance D is calculated by

$$D = \frac{k_1}{k_1 + k_2}d_1 + \frac{k_2}{k_1 + k_2}d_2 \tag{10}$$

where k_1 and k_2 are the numbers of the samples of class C_1 and class C_2. The weighted inner distance D can consider the distances within class C_1 and C_2 in an integrated manner.

(5) Feature selection index J_f. According to the J and D calculated previously, the new feature selection index can be obtained by

$$J_f = \frac{J}{D} \tag{11}$$

This formula indicates that the larger the J is and the smaller the D is, the bigger the J_f is. In this condition, the separability of the feature is better and vice versa. Therefore, the features with larger J_f are more significant for classification. However, for the better portability of the method, the selected features should not be too many. In this paper, the first two features of each two classes are selected as the optimum features for classification. The overall classification process is shown in Fig.1.

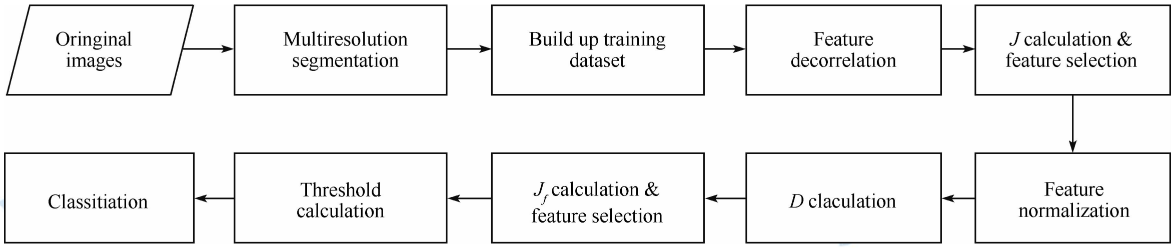


Fig.1 The overall classification flow chart

3 EXPERIMENT AND ANALYSIS

3.1 Test area and data

The improved method is compared with the SEaTH algorithm by performing object-oriented classification of the HJ and TM images of Zhaoqing city in Guangdong province (Fig.2). Zhaoqing is a model of the national environmental protection city. It is characterized by subtropical monsoon climate with four distinct seasons. The vegetation in the area is rich and the terrain is complex. Numerous meteorological disasters such as rainstorm, flood, high temperature, and drought occur frequently. Therefore, studying the distribution of the land cover type of Zhaoqing has great significance for its environmental protection. The information on experimental data and verifying data is presented in Table 1, and the distribution of the verification point is reported in Fig.3. The eCognition software is used in the sample selection and classification.

Table 1 The information of the test data

No.	Sensor	Orbit	Date	Resolution/m
1	HJ-1B CCD	1—88	March 11, 2010	30
2	HJ-1B CCD	1—89	December 4, 2010	30
3	Landsat 7 ETM+	123—44	September 1, 2010	30
4	Others	ASTER GDEM		30
5		LBV transformation images of data 2		30
6	Verifying data	Ground sampling points and high-resolution images from Google Earth		

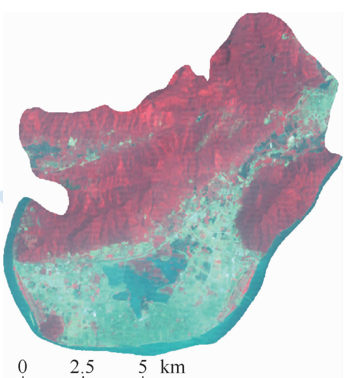


Fig.2 The geographical position schematic diagram of the test area

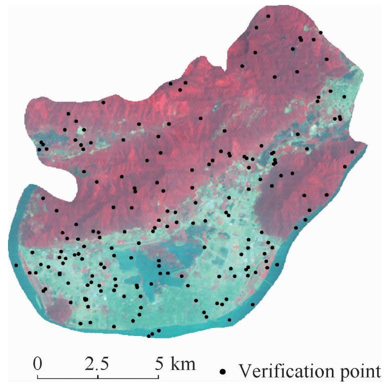


Fig.3 The distribution graph of the verification point

3.2 Data preprocessing

The data preprocessing in this study mainly involves geometric correction, radiation calibration, strip removement, topographic correction, and LBV transformation.

Geometric correction: ETM + L1T image is already corrected precisely and HJ-1A/1B L1 image has coordinate information. However, large deformation is observed on the edge of the HJ image, so the registration should be done between the HJ and ETM+ images. In this study, the quadratic polynomial correction model and the adjacent resampling method of the ENVI software perform this work.

Radiation calibration: The radiation calibration of the ETM + image is conducted with the Landsat calibration module of the ENVI software, and the HJ images are calibrated based on the coefficients published by the China Resources Satellite Application Center.

Strip removement: This study is mainly focused on the banding loss problem of ETM+ images.

Topographic correction: The images are corrected by ASTER GDEM.

LBV transformation: According to the LBV transformation method (Zeng, 2007), the transformation formula for the HJ multispectral images is deduced and given by

$$L = \hat{D}_{0.7} = -0.2005D_1 + 0.4055D_2 + 0.7358D_3 + 0.1183D_4 \tag{12}$$

$$V = v_1 - v_2 + v_3 - v_4 = -0.5575D_1 + 1.3338D_2 - 0.8967D_3 + 0.1203D_4 \tag{13}$$

$$B = -b = 1.9571D_1 + 0.9630D_2 - 0.1553D_3 - 2.7648D_4 \tag{14}$$

The LBV transformation is efficient in extracting vegetation, bodies of water, and artificial surfaces, so it is used in the classification. The reflection of the images is expanded 10,000 times before being segmented because the eCognition software cannot segment the images with float values. Then, an extremely simple segmentation procedure is applied with scale factor of 30, shape factor of 0.1, and compactness of 0.5.

In this study, the images are classified into five classes: forest, cropland, wetland, artificial surface, and others.

3.3 Experiment results and analysis

To compare the application effect of the SEaTH algorithm and the improvement method in land cover classification, the features are selected via three methods: (1) the original SEaTH algorithm, (2) the SEaTH algorithm with decorrelation, and (3) the improved method proposed in this paper. The segmentation results in 4343 image objects and 109 samples are selected, which provide 215 original features (Table 2). The results of the feature selection and threshold calculation are reported in Table 3.

Table 2 The original feature set

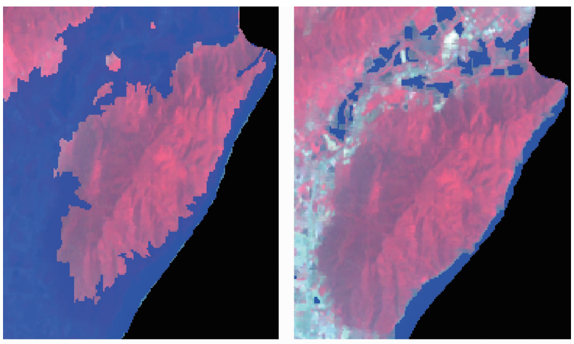
Spectral features	Shape features	Texture features
Mean	Area	GLCM Homogeneity
Standard deviation	Border length	GLCM Contrast
Maximum differential	Length	GLCM Dissimilarity
Brightness	Width	GLCM Entropy
NDVI	Length/width	GLCM Ang. 2nd moment
NDWI	Asymmetry	GLCM Mean
BAI	Compactness	GLCM StdDev
RRI	Density	GLCM Correlation
NDGI	Elliptic fit	
SBI	Main direction	
	Rectangular fit	
	Shape index	

Table 3 The results of feature selection and threshold calculation

Class	Method	Feature	J	J_f	Omen	Threshold
Forest from Artificial surface	The first	NDVI12	1.99997117	10.27196681	>	0.42438
		Mean_TM0901_3	1.99966058	19.28704111	<	860.15310
	The second	Mean_TM0901_3	1.99966058	19.28704111	<	860.15310
		RRI12	1.99708628	11.13284742	<	0.33872
	The third	Mean_TM0901_3	1.99966058	19.28704111	<	860.15310
		Mean_DEM	1.92599969	16.99376047	>	25.82830
Forest from Wetland	The first	NDVI12	1.92696023	9.59152978	>	0.54222
		NDWI12	1.91646396	9.26461049	<	−0.50348
	The second	Mean_TM0901_5	1.86643553	12.96587952	>	682.17620
		Mean_TM0901_4	1.77799689	10.31534917	>	964.98340
	The third	GLCM_Mean_DEM	1.45393791	18.19258302	>	125.08490
		GLCM_Ang_2_DEM	1.76954863	17.60126486	<	0.006631
Artificial surface from Cropland	The first	Mean_TM0901_4	1.93137638	11.07554497	<	1435.50010
		Max_diff	1.90015117	14.71414946	<	0.94229
	The second	Mean_TM0901_4	1.93137638	11.07554497	<	1435.50010
		Max_diff	1.90015117	14.71414946	<	0.94229
	The third	Max_diff	1.90015117	14.71414946	<	0.94229
		RRI12	1.68712599	12.50484458	>	0.46765

Note: Table 3 only shows a part of the typical results

Table 3 implies the following conclusions: (1) The decorrelation analysis can select the features with small correlation to ensure that most information is used in the classification, and that information redundancy is avoided. For example, in distinguishing the artificial surface and forest, NDVI and the mean value of TM band 3 are selected by the first algorithm, whereas the ratio resident-area index (RRI) (Wu, et al., 2006) is selected instead of NDVI by the second algorithm. The identification of the two indices suggests that the features selected by the second method contain more information than those selected by the first method. (2) The new index J_f is better than J in selecting the classification features. For example, in distinguishing forest and wetland, the extraction result of the second method is better than that of the first (Fig.4). Fig.4 sugg-

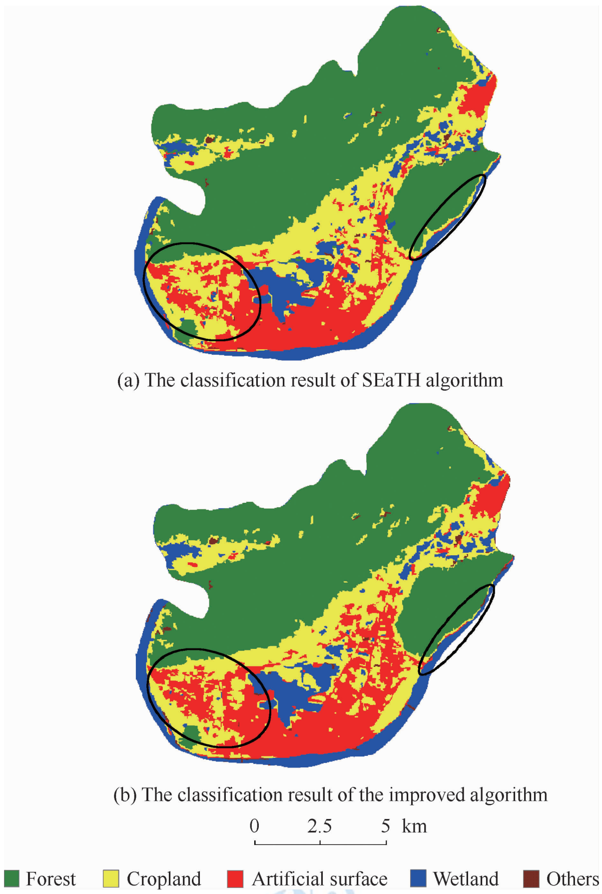


(a) Wetland extraction result of the first method (b) Wetland extraction result of the second method

Fig.4 Wetland extraction results of two methods ests that it may be better for the first two features to be used alone because the information they contain is similar and so are the extracted objects, whereas the latter two features containing richer information can have a complementary function in extracting objects more accurately. Meanwhile, feature decorrelation can remove the redundancy and not the useful information. (3) The features filtered out by the first method are concentrated in the mean values of spectral bands, whereas the third method filtered out more diversified features such as maximum differential (Max _ diff), brightness, standard deviation, and texture features. (4) The result of the first and the third methods in distinguishing cropland and artificial surface suggests not only that RRI is efficient in extracting artificial surfaces, but also that third method can filter out better features than the first one. (5) These results can be used as a reference for other classification projects. (6) Throughout the feature selection results in Table 3, regardless of which method is used, no shape features are observed, which indicates that for low-resolution remote sensing images, the separability of the shape features is not stronger than that of the spectral and texture features.

The classification results of the first and the third methods are illustrated in Fig.5, and their accuracy assessments are also presented in Table 4 and Table 5.

The classification results indicate that forest and artificial surfaces can easily be mixed with cropland at their edges by the first method. The accurate distinction of the three features has been a difficulty of classification by a single image, but the proposed method improves the classification results to a certain extent. Specifically, the accuracy of cropland is increased by



(a) The classification result of SEaTH algorithm

(b) The classification result of the improved algorithm

Fig.5 Two classification maps

Table 4 The confusion matrix of the first result

Sample data		Reference data						
		Forest	Cropland	Artificial surface	Wetland	Others	Sum	User accuracy/%
Results	Forest	53	1	0	1	1	56	94.64
	Cropland	7	28	16	5	0	56	50.00
	Artificial surface	0	5	50	2	0	57	87.72
	Wetland	0	0	0	20	0	20	100
	Others	0	0	0	0	1	1	100
	Sum	60	34	66	28	2	190	
	Producer accuracy/%	88.33	82.35	75.76	71.43	50.00		
	Overall accuracy = 80.00% Kappa = 0.7277							

Table 5 The confusion matrix of the second result

Sample data		Reference data						
		Forest	Cropland	Artificial surface	Wetland	Others	Sum	User accuracy/%
Results	Forest	53	0	1	0	1	55	96.36
	Cropland	7	33	11	2	0	53	62.26
	Artificial surface	0	1	54	5	0	60	90.00
	Wetland	0	0	0	21	0	21	100
	Others	0	0	0	0	1	1	100
	Sum	60	34	66	28	2	190	
	Producer accuracy/%	88.33	97.06	81.82	75.00	50.00		
	Overall accuracy = 85.26% Kappa = 0.7989							

12.26% because the latter classification features are richer, so that the classes can be separated in different aspects.

However, compared with other categories, the classification accuracy of the cropland remains low, and it is mixed with artificial surface to a larger extent. The main reason is that the images are acquired in September and December when most of the cropland in Zhaoqing is in the harvest period, thus reflecting the bare surface characteristics, which can be mixed easily with the construction land. The classification accuracies of the four other features are more than 90% and the overall accuracy is as much as 85.26%, which basically meets the requirements of automatic classification accuracy. Generally, the improved method is more efficient than the original SEaTH algorithm.

4 CONCLUSIONS

In this study, an improved feature selection method is proposed and applied to object-oriented land cover classification. By comparing and analyzing the experiment results, this study obtains several conclusions below.

(1) Decorrelation can increase the effective information from the selected features and prevent information redundancy.

(2) The improved method selects more diversified features such as texture features, maximum differential, and brightness, except for spectral features.

(3) For low-resolution remote sensing images, the shape features are not extremely effective in distinguishing surface features.

(4) The improved method can effectively improve the classification accuracy of the artificial surface, forest, and cropland, which are difficult to separate via remote sensing images, particularly the multi-temporal images without good quality. Thus, this method is useful in improving the single-temporal image classification result.

However, this proposed method has several limitations: (1) This approach is more effective for the rough categories, but requires further validation for the more detailed categories (such as woodland leaved forest and coniferous forest). (2) The classification accuracy of the cropland in this experiment is not high enough, and requires further research to improve its classification accuracy. In addition, the influence of the classification order on its result also requires investigation.

REFERENCES

Addink E A, de Jong S M, Davis S A, Dubyanskiy V, Burdelow L A and Leirs H. 2010. The use of high-resolution remote sensing for plague surveillance in Kazakhstan. Remote Sensing of Environment, 114 (3): 674−681 [DOI: 10.1016/j.rse.2009.11.015]

Bruzzone L and Serpico S B. 2000. A technique for feature selection in multiclass problems. International Journal of Remote Sensing, 21(3): 549−563 [DOI: 10.1080/014311600210740]

Chen J, Deng M, Xiao P F, Yang M H and Mei X M. 2010. Rough set theory based object-oriented classification of high resolution remotely sensed imagery. Journal of Remote Sensing, 14(6): 1139−1155

Feng L, Li M H and Li F X. 2008. Texture feature selection in remote sensing image based on genetic algorithms. Journal of Nanjing University (Natural Sciences), 44(3): 310−319

Gao W. 2010. The Study of Information Extraction Technology for Remote Sensing Based on Feature Knowledge. Beijing: China University of Geosciences

Gao Y, Marpu P, Niemeyer I, Runfola D M, Giner N M, Hamill T and Pontius R G Jr. 2011. Object-based classification with features extracted by a semi-automatic feature extraction algorithm-SEaTH. Geocarto International, 26(3): 211−226 [DOI: 10.1080/10106049.2011.556754]

Han P, Gong J Y, Li Z L, Bo Y C and Cheng L. 2010. Selection of optimal scale in remotely sensed image classification. Journal of Remote Sensing, 14(3): 507−518

He X C and Xu S S. 2002. Method of feature selection based on association rules. Infrared and Laser Engineering, 31(6): 504−507

Ji X J, Li S Z and Li T. 2001. Application of the correlation analysis in feature selection. Journal of TEST and Measurement Technology, 15 (1): 15−18

Koller D and Sahami M. 1996. Toward Optimal Feature Selection // Proceedings of International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann: 284−292

Laliberte A S, Browning D M and Rango A. 2012. A comparison of three feature selection methods for object-based classification of sub-decimeter resolution UltraCam-L imagery. International Journal of Applied Earth Observation and Geoinformation, 15: 70 − 78 [DOI: 10.1016/j.jag.2011.05.011]

Nussbaum S, Niemeyer I and Canty M J. 2006. SEaTH-a new tool for automated feature extraction in the context of object-based image analysis // Proceedings of the 1st International Conference on Object-based Image Analysis. Austria: Salzburg University

Peng H, Long F and Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8): 1226−1238 [DOI: 10.1109/TPAMI.2005.159]

Sun J X. 2002. Modern Pattern Recognition. Beijing: Higher Education Press

Wu B, Zhu Q D, Gao H Y and Zhou X C. 2009. Feature selection based on maximal mutual information criterion in object-oriented classification. Remote Sensing for Land and Resources, 81(3): 30−34

Wu H A, Jiang J J, Zhang H L, Zhang L and Zhou J. 2006. Application of ratio resident-area index to retrieve urban residential areas based on Landsat TM data. Journal of Nanjing Normal University (Natural Science), 29(3): 118−121

Ye Z W, Zheng Z B, Wan Y C and Yu X. 2007. A novel approach for feature selection based on ant colony optimization algorithm. Geomatics and Information Science of Wuhan University, 32(12): 1127−1130

Zeng Z Y. 2007. A new method of data transformation for satellite images: I. Methodology and transformation equations for TM images. International Journal of Remote Sensing, 28(18): 4095−4123 [DOI: 10.1080/01431160601028912]

Zhang X Y, Feng X Z, and Jiang H. 2009. Feature set optimization in object-oriented methodology. Journal of Remote Sensing, 13 (4): 664−669

Zhang Z P. 2006. A Feature Selection Algorithm Based on Minimum Upper Bound of Bayes Error. Shanghai: Tongji University

Zheng Y. 2010. Study of Extraction Information of Land Cover Based on SPOT5 Panchromatic Images. Beijing: China University of Geosciences

面向对象分类特征优化选取方法及其应用

王贺，陈劲松，余晓敏

中国科学院 深圳先进技术研究院,广东 深圳 518055

摘　要:分离阈值算法 SEaTH 是一种有效的自动选取分类特征并计算阈值的方法,但其只考虑了类间距离,容易存在信息的冗余,从而对分类精度造成一定影响。本文在 SEaTH 算法的基础上,综合考虑了特征间的相关性、类间距离以及类内距离,对 SEaTH 算法进行了优化,并将改进前后的两种方法运用到广东省肇庆市 TM 影像及环境一号卫星影像土地覆盖分类中进行对比分析。实验结果表明,改进后的方法筛选出的特征在提取地物上更为有效,尤其使耕地的分类精度提高了 12.26%,使分类总体精度由 80% 提高到 85.26%。耕地与林地分类精度的提高,对不易获取质量较好的多时相数据地区的土地覆盖分类具有重要意义。

关键词:面向对象分类,特征筛选,SEaTH 算法,土地覆盖分类,特征去相关

中图分类号:P237　　**文献标志码:**A

引用格式:王贺,陈劲松,余晓敏. 2013. 面向对象分类特征优化选取方法及其应用. 遥感学报,17(4): 816–829

Wang H, Chen J S and Yu X M. 2013. Feature selection and its application in object-oriented classification. Journal of Remote Sensing, 17(4): 816–829 [DOI: 10.11834/jrs.20132257]

1 引　言

土地覆盖分类是遥感技术的重要应用之一,也是研究区域气候、生态和环境等基本状况的重要基础。如今,随着社会经济的发展与科技的进步,社会问题也日趋突出,如全球气候变化、生态环境遭到破坏等,给人类的生存、发展带来了严重后果。如何能够快速、准确、及时地掌握地表覆盖信息,从而为生态环境保护以及经济发展规划等提供可靠的基础数据,是急需解决的问题。另外,如何选取出有效的分类特征,提高分类精度并增加可分的地物类型,长久以来都是遥感土地覆盖分类急需突破的一个关键问题。与此同时,随着研究问题的不断深入,更加高效、精确、细致地获取地表信息也对遥感土地覆盖分类方法及其关键技术提出了更高的要求。

可用于分类的光学遥感影像特征多种多样,最基本的如光谱特征、纹理特征和几何特征。随着遥感影像数据源的增多以及学者对遥感影像分类方法的深入研究,多源遥感数据的各种基本特征被进一步演变和利用,演绎出更多的可用信息。与传统基于像元的分类方法相比,面向对象分类方法不仅可以利用光谱特征和纹理特征,还可以将几何特征和上下文信息加入到分类特征中,从多种角度挖掘地物的特性,从而提高地物的可分性及分类精度(Laliberte 等,2012)。然而,特征种类、维数的增多使得快速、高效地选取出有效分类特征成为难点,并且针对不同地物特征合理选取阈值也并不容易。为了解决上述问题,一些分类的参数选取方法相继被提出(Koller 和 Sahami,1996;冯莉 等,2008;高伟,2010;Bruzzone 和 Serpico,2000;张振平,2006;叶志伟 等,2007;何小晨和徐守时,2002;Peng 等,2005;Nussbaum 等,2006;郑毅,2010;Addink 等,2010;陈杰 等,2010;韩鹏 等,2010;吴波 等,2009),其中,分离阈值法 SEaTH 是目前比较有代表性的面向对象的特征选取方法。该方法是由 Nussbaum 等人(2006)提出的一种基于 Jeffries- Matudita 距离和高斯分布混合模型的分类特征选取与阈值计算方法,已有研究表明,该方法人工干预较少且执行效率较高,与最近邻分类法和最大似然分类法相比,其分

收稿日期:2012-09-11;**修订日期:**2013-01-08;**优先数字出版日期:**2013-01-15

基金项目:中国科学院战略性先导科技专项(编号:XDA05050107-03)

第一作者简介:王贺(1986—),女,硕士,助理工程师,主要从事多源遥感影像信息提取与应用研究。E-mail: he.wang@siat.ac.cn

通信作者简介:余晓敏(1982—),女,博士,助理研究员,主要从事遥感动态变化检测研究。E-mail: yu.x.min@gmail.com

类效果更好(Gao 等,2011)。然而,SEaTH 算法只考虑了各类别之间的概率距离,并没有充分考虑到特征之间的相关性以及各类别内部的差异性,筛选出的特征易出现信息冗余,从而影响分类效果。

为解决以上问题,本文在 SEaTH 算法的基础上,对所有分类特征之间的相关性进行分析,综合考虑了类间距离与类内距离,以类间距离为主、类内距离为辅的原则来筛选分类特征,以此获得最优分类特征子集。然后利用面向对象的分类方法,以中分辨率的 Landsat TM 影像以及中国环境卫星(HJ-1A/1B)多光谱影像为实验数据,以广东省肇庆市为实验区进行土地覆盖分类研究。

2 原理与方法

2.1 SEaTH 算法

SEaTH 算法(Nussbaum 等,2006)是一种基于训练样本的统计方法,其主要步骤如下:

(1)选取样本。建立包含所有目标地物类型的训练样本区(样本个数约占所有影像对象的 2%),并输出所有样本的待选特征。

(2)筛选最优分类特征。对于一个给定的训练样本区,估计该样本区内每一类地物的概率分布,从而计算给定的两类地物 C_1 和 C_2 之间的 Jeffries-Matusita 距离 J 作为这两类地物间的可分度,具体公式如下:

$$J = 2(1 - e^{-B}) \tag{1}$$

$$B = \frac{1}{8}(m_1 - m_2)^2 \frac{2}{\sigma_1^2 + \sigma_2^2} + \frac{1}{2}\ln\left(\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}\right) \tag{2}$$

式中,m_i 和 σ_i^2, $i=1,2$ 分别代表两个地物类型对应给定特征的均值和方差。由式(1)可知,J 取值范围为[0,2]。当 J 值趋向于 0 时,特征的可分度较小;当 J 值趋向于 2 时,特征的可分度较大。因此,选出 J 值较大的特征,组成用于地物分类的最优特征子集。

(3)计算各特征的分类阈值。当给定的特征符合正态概率分布时,该特征的任意值 x 满足混合高斯概率分布模型,则对于分类阈值,应该满足

$$p(x \mid C_1)P(C_1) = p(x \mid C_2)p(C_2) \tag{3}$$

式中,$p(x|C_1)$ 是一个均值为 m_{c1}、方差为 σ_{c1}^2 的标准正态分布概率函数,$p(x|C_2)$ 是一个均值为 m_{c2}、方差为 σ_{c2}^2 的标准正态分布概率函数。计算该特征分类阈值的具体公式如下:

$$x_{1(2)} = \frac{1}{\sigma_{c1}^2 - \sigma_{c2}^2}\left[m_{c2}\sigma_{c1}^2 - m_{c1}\sigma_{c2}^2 \pm \sigma_{c1}\sigma_{c2}\sqrt{(m_{c1} - m_{c2})^2 + 2A(\sigma_{c1}^2 - \sigma_{c2}^2)}\right] \tag{4}$$

$$A = \log\left[\frac{\sigma_{c1}}{\sigma_{c2}} \times \frac{p(C_2)}{p(C_1)}\right] \tag{5}$$

由式(4)和式(5),计算出两个分类阈值 x_1 和 x_2,选取大小位于 m_{c1} 和 m_{c2} 之间的阈值作为影像分类的阈值。

2.2 改进后的特征优化选取算法

面向对象分类主要基于分割后的影像对象的光谱、纹理和形状等众多特征进行分类。这些特征之间常常存在着很大的相关性,造成信息的冗余。已有理论证明:在原特征集中增加或删除相关的特征,不影响该特征的分类能力(吉小军 等,2001)。因此,本文在 SEaTH 算法的基础上引入了特征之间的去相关处理,先初步去除相关性较大的特征,然后综合考虑类间距离与类内距离的共同影响,以类间距离为主、类内距离为辅的原则构建了新的特征筛选指标。改进后的特征筛选步骤如下:

(1)计算原始特征集的自相关矩阵,根据特定的阈值,删除相关性较大的若干特征,即特征的去相关处理。设原始高维特征集为 $\boldsymbol{F}_N = (f_1, f_2, \cdots, f_N)$,样本对象个数为 K,则 N 个特征的自相关系数矩阵可表示为:

$$\boldsymbol{r} = \begin{pmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & & \vdots \\ r_{N1} & \cdots & r_{NN} \end{pmatrix} \tag{6}$$

$$r_{ij} = \frac{\sum_{l=1}^{K}(x_i^l - \bar{x}_i)(x_j^l - \bar{x}_j)}{\sqrt{\sum_{l=1}^{K}(x_i^l - \bar{x}_i)^2 \sum_{l=1}^{K}(x_j^l - \bar{x}_j)^2}} \quad (i = 1, \cdots, N; j = 1, \cdots, N) \tag{7}$$

式中,x_i^l 表示特征集中第 l 个样本第 i 个特征的值,$\bar{x}_i$ 表示第 i 个特征的均值估计,$\bar{x}_i = \frac{1}{K}\sum_{l=1}^{K} x_i^l$。

计算出每两个特征间的相关系数后,根据设定的阈值和图像信息量的度量准则:方差越小,特征包含的分类信息越少;方差越大,特征包含的信息越丰富(张秀英 等,2009),去除相关性较大的特征对中方差较小的一个特征。设置不同的阈值可调节入选特征的数目,阈值越小,入选特征个数越少,特征间的相关性也越小,本文经多次试验后设定相关系数阈值为 0.95。去相关后的特征集 $\boldsymbol{F}_M(M<N)$ 中

既包含所有原始特征中的大部分信息,又避免了信息的冗余,可作为下一步筛选特征的依据。

(2)计算类间距离。根据式(1)计算特征集 $\boldsymbol{F}_M$ 中所有类别两两之间的类间距离 J,并对其进行降序排列,选取出每对类别中类间距离最大的前 10 个特征,构成特征集 $\boldsymbol{F}_p$ 作为下一步筛选特征的依据。

(3)特征归一化处理。由于所选特征数量级不同,为了便于不同特征之间的计算与比较,在计算类内距离之前,需对上一步中选出的 10 个特征进行归一化处理,使其取值范围都位于 0 至 1 之间,具体公式如下:

$$F' = \frac{F - F_{\min}}{F_{\max} - F_{\min}} \tag{8}$$

(4)计算每一类的类内距离,并加权平均。设归一化后的特征集为 $\boldsymbol{F}_p' = (f_1, f_2, \cdots, f_p)$,对象类别为 $\boldsymbol{C}_n = (c_1, c_2, \cdots, c_n)$,从各类别中选取的典型样本个数为 $\boldsymbol{K}_n = (k_1, k_2, \cdots, k_n)$,则类内距离为

$$d = \frac{2}{k(k-1)} \sum_{l=1}^{k} \sum_{m=l+1}^{k} (x_l - x_m)^2 \tag{9}$$

式中,x_l 表示某类第 l 个样本给定特征的值,x_m 表示某类第 m 个样本给定特征的值,k 表示样本数。依次遍历类别 C_1 和 C_2 中的每个样本,计算每个样本与同类其他样本的某个特征的值(如 $f_j, j=1,\cdots,p$)的距离并累加,分别得到 C_1 和 C_2 的类内距离 d_1 和 d_2(孙即祥,2002),再根据 C_1 和 C_2 类的样本数分别赋予类内距离 d_1 和 d_2 相应的权重,得到加权类内距离

$$D = \frac{k_1}{k_1 + k_2} d_1 + \frac{k_2}{k_1 + k_2} d_2 \tag{10}$$

式中,k_1 和 k_2 分别表示 C_1 和 C_2 类的样本数。用加权类内距离 D 能够综合考虑类别 C_1 和 C_2 的类内距离。

(5)构建特征筛选指标 J_f。根据上面计算出的类间距离 J 和加权类内距离 D,可以得到新的特征筛选指标:

$$J_f = \frac{J}{D} \tag{11}$$

从 J_f 的构造可以看出,特征的类间距离越大、类内距离越小,则 J_f 的值越大,说明特征对不同类别的分离性越好;反之,J_f 的值越小,说明特征的分离性越差。对 J_f 进行降序排列,排在前面的若干特征对分类意义较大,而为使方法移植效果更好,入选的特征不能太多。考虑到 eCognition 软件操作的方便性,本文选取前两个特征作为最后的分类特征,并根据式(5)计算各特征的分类阈值。总体分类流程如图 1 所示。

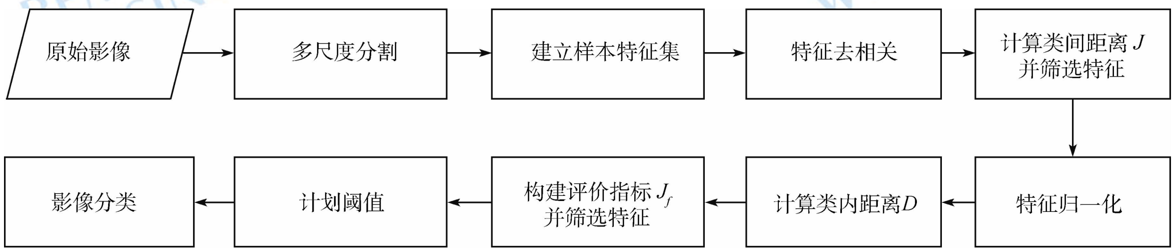


图 1　总体分类流程图

3　实验结果与分析

3.1　实验区及实验数据

实验区位于广东省肇庆市境内(图 2),地理位置为(111°3′E—112°53′E,22°23′N—24°23′N)。肇庆市属南亚热带季风气候,四季特点较为分明,植被类型丰富,地形复杂,北部、西部及南部与外县交界处是山地,中部和南部大部分地区为丘陵,东南为平原。肇庆境内较常发生暴雨、洪涝、高温和寒旱等气象灾害。因此,研究肇庆市土地覆盖类型的分布,对其生态环境保护具有重要的意义。本文的实验数据与验证数据如表 1 所示,验证点分布情况如图 3 所示。另外,本文中的样本选取以及分类工作是在 eCognition 软件中实现的。

表 1　实验数据信息表

序号	数据类型	轨道号	获取日期	地面分辨率/m
1	HJ-1B CCD	1—88	2010-03-11	30
2	HJ-1B CCD	1—89	2010-12-04	30
3	Landsat 7 ETM+	123—44	2010-09-01	30
4	其他	ASTER GDEM		30
5	其他	影像 2 变换后的 L、B、V 图像		30
6	验证数据	地面采样点及 Google Earth 高分辨率影像		

3.2　数据预处理

数据预处理主要包括卫星影像的几何校正、辐

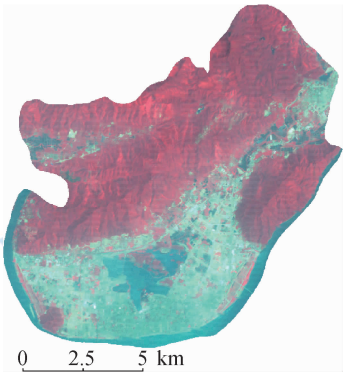


图 2　实验区地理位置示意图

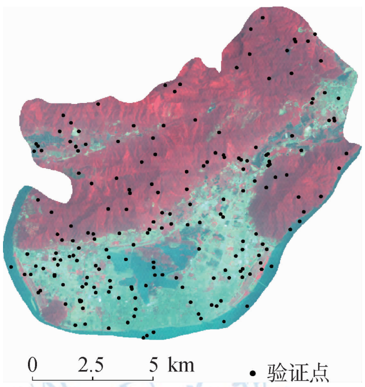


图 3　验证点分布图
(R: Band_4; G: Band_3; B: Band_2)

射定标、去条带处理、地形校正以及 LBV 变换等。

几何校正：ETM+ L1T 影像已经做过精确的几何校正，HJ-1A/1B 一级数据本身也带有坐标，但由于 HJ-1A/1B 影像的幅宽较宽，其星下点误差较小，周边的变形较大，因此需要对 HJ 卫星影像和 ETM+ 影像进行图像配准。本文采用 ENVI 软件中的二次多项式纠正模型和最邻近重采样方法进行图像的配准。

辐射定标：影像的辐射定标也是在 ENVI 软件中进行的。其中 ETM+影像采用 ENVI 软件中自带的 Landsat Calibration 模块实现，而 HJ 卫星影像根据中国资源卫星应用中心公布的定标系数进行影像计算实现。

去条带：由于 2010 年 ETM+影像上有数据条带丢失，因此在分类前还对影像进行了去条带处理。

地形校正：利用 30 m 分辨率的 ASTER GDEM 对影像进行了地形校正。

LBV 变换：本文在分类时不仅运用了原始的环境卫星影像，还对其进行了 LBV 变换(Zeng，2007)。根据 Zeng(2007)提出的 LBV 变化方法，本文在分类前首先对中国环境卫星多光谱影像的 LBV 变换公式进行了推导，得到了适用于 HJ-1B 卫星多光谱影像的 LBV 变换公式：

$$L = \hat{D}_{0.7} = -0.2005D_1 + 0.4055D_2 + 0.7358D_3 + 0.1183D_4 \quad (12)$$

$$V = v_1 - v_2 + v_3 - v_4 = -0.5575D_1 + 1.3338D_2 - 0.8967D_3 + 0.1203D_4 \quad (13)$$

$$B = -b = 1.9571D_1 + 0.9630D_2 - 0.1553D_3 - 2.7648D_4 \quad (14)$$

LBV 变换对植被、水体及人工表面具有很好的提取效果，因此将变换后的 LBV 图像加入到分类数据中，共同参与土地覆盖分类。

另外，由于 eCognition 软件不能对浮点数影像进行分割，因此在分割前还要将影像的反射率扩大 10000 倍，并将分割尺度设为 30，形状系数为 0.1，紧致度为 0.5。

本文将地物类型分为 6 个大类，即林地、耕地、草地、湿地、人工表面以及其他类。其中，其他类包括裸土、裸岩、稀疏植被和盐碱地等。由于实验区内没有典型的草地样本，本文实验只对其余 5 个大类进行区分。

3.3　实验结果与分析

为了比较分析改进 SEaTH 算法在土地覆盖分类中的应用效果，分别基于 3 种方法进行特征筛选与阈值计算：(1)原始 SEaTH 算法；(2)加入去相关的 SEaTH 算法；(3)本文提出的改进 SEaTH 算法。本文共分割 4343 个对象，对林地、耕地、湿地、人工表面以及其他类共选取样本 109 个，得到原始特征共 215 个(表 2)。3 种方法特征筛选及其阈值计算结果如表 3 所示。

表 2　原始特征集

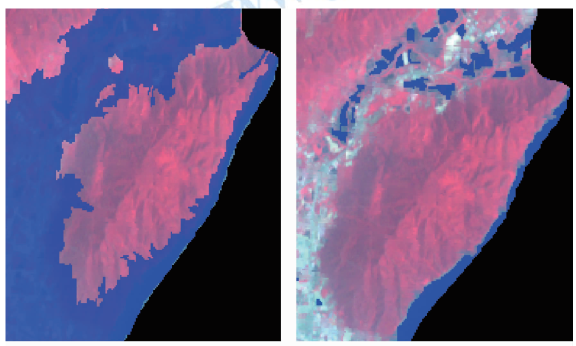
光谱特征	形状特征	纹理特征
	面积	
均值	边界长度	
标准差	长度	GLCM 均质性
最大差分	宽度	GLCM 对比度
亮度	长宽比	GLCM 相异性
归一化植被指数	不对称性	GLCM 熵
归一化水体指数	紧凑度	GLCM 角二阶矩
建成区面积指数	密度	GLCM 均值
比值居民地指数	椭圆符合度	GLCM 标准差
归一化差异绿度指数	主方向	GLCM 相关性
土壤亮度指数	矩形符合度	
	形状指数	

表3 3种特征筛选方法结果对比

地物类型	特征筛选方法	分类特征	J	J_f	数量关系	分类阈值
林地（与人工表面区分）	原始 SEaTH 算法	NDVI12	1.99997117	10.27196681	>	0.42438
		Mean_TM0901_3	1.99966058	19.28704111	<	860.153100
	加入去相关后的 SEaTH 算法	Mean_TM0901_3	1.99966058	19.28704111	<	860.153100
		RRI12	1.99708628	11.13284742	<	0.33872
	考虑类内距离的 SEaTH 算法	Mean_TM0901_3	1.99966058	19.28704111	<	860.153100
		Mean_DEM	1.92599969	16.99376047	>	25.828300
林地（与湿地区分）	原始 SEaTH 算法	NDVI12	1.92696023	9.59152978	>	0.54222
		NDWI12	1.91646396	9.26461049	<	-0.50348
	加入去相关后的 SEaTH 算法	Mean_TM0901_5	1.86643553	12.96587952	>	682.176200
		Mean_TM0901_4	1.77799689	10.31534917	>	964.983400
	考虑类内距离的 SEaTH 算法	GLCM_Mean_DEM	1.45393791	18.19258302	>	125.084900
		GLCM_Ang_2_DEM	1.76954863	17.60126486	<	0.006631
人工表面（与耕地区分）	原始 SEaTH 算法	Mean_TM0901_4	1.93137638	11.07554497	<	1435.500100
		Max_diff	1.90015117	14.71414946	<	0.942290
	加入去相关后的 SEaTH 算法	Mean_TM0901_4	1.93137638	11.07554497	<	1435.500100
		Max_diff	1.90015117	14.71414946	<	0.94229
	考虑类内距离的 SEaTH 算法	Max_diff	1.90015117	14.71414946	<	0.94229
		RRI12	1.68712599	12.50484458	>	0.467650

注：表3中只列出部分具有代表性的特征筛选结果

分析表3可以得到以下结论：(1)加入去相关分析后，确实可以保留相关性小的特征。例如，在区分人工表面和林地时，原始 SEaTH 算法筛选出的特征是 NDVI 以及 TM 第3波段均值，由 NDVI 的公式可以知道两者之间具有很大的相关性，存在着信息冗余，而加入相关分析的 SEaTH 算法则将 NDVI 排除，换成了比值居民地指数 RRI(吴宏安 等，2006)，从而保证了最大限度地应用原始特征集中包含的信息，同时避免了信息冗余。(2)虽然进行去相关分析后可能将一些类间距离大的特征排除，但是补充上的特征能够更好地提取出第一特征提取不出的对象。例如在从林地中提取湿地时，原始 SEaTH 算法筛选出的特征分别是 NDVI 和 NDWI，二者的类间距离分别为 1.92696023 和 1.91646396，而加入去相关分析后筛选出的两个特征 Mean_TM5 和 Mean_TM4 的类间距离分别为 1.86643553 和 1.77799689，小于前两个特征的类间距离，但是后者的 J_f 指标值大于前者，并且其分类结果好于前两个特征(图4)。



(a) 原始算法湿地提取结果　(b) 去相关后湿地提取结果

图4 两种不同方法提取水体结果对比

由图4可以看出，在不考虑其他类别而只考虑林地和湿地两类区分结果的情况下，原始 SEaTH 算法在林地和湿地边缘处有很大的混淆，而加入去相关分析后湿地提取效果明显好于原始 SEaTH 算法。这一方面说明，两组特征在提取地物时前者也许在单独使用时效果较好，但是由于两个特征包含的信息相似，提取的对象重复性较大；而后者中两个特征包含信息更丰富，可以起到相互补充的作用，能够更精确地提取地物。另一方面说明，去相关并不会将真正有用的信息去除，而只是将冗余的信息去除。综上所述，加入相关分析是必要的，可以将更好的特征筛选出来。(3)原始 SEaTH 算法筛选出的特征大部分集中于各波段光谱的均值，而改进后的 SEaTH 算法筛选出的特征更加多元化，不仅仅集中

于光谱均值，还加入了最大差分(Max_diff)、亮度(Brightness)、方差(Stddev)以及纹理信息(GLCM)等特征，更有利于地物的区分。(4)在区分人工表面和耕地时，原始 SEaTH 算法筛选出的特征为 TM4 和 Max_diff，而改进后的 SEaTH 算法筛选出的特征为 RRI 和 Max_diff，这不仅说明了比值居民地指数在提取人工地物时的有效性，还说明了改进后的 SEaTH 算法能够将更优的特征筛选出来。(5)对具体类别进行分析可以得到提取每个类别的最好特征，如对于人工表面来说，区分其与其余类别的最优特征是 Max_diff，而提取林地的最优特征则是 DEM，提取其他类别的最优特征是 TM3 的方差等。虽然这一结论目前只适用于本实验中的数据，但是也对今后的土地覆盖分类特征选取起到了一定的提示作用。(6)纵观表 3 中 3 种方法的特征筛选结果可以发现，无论用哪种方法，选出的特征都不包含形状特征，这也说明对于中低分辨率的遥感影像来说，形状特征对不同地物的区分能力没有光谱特征和纹理特征强。

另外，本文提出的特征筛选方法是以类间距离为主、类内距离为辅的宗旨来筛选特征的，即先根据计算出的类间距离对所有特征进行排序，然后计算前十个特征的类间距离与类内距离的比值，再对比值进行降序排列，选取前两个对应的特征作为最终的分类特征，这样就避免了类间距离与类内距离都较小，但它们的比值很大的情况发生。

在 eCognition 软件中分别利用原始 SEaTH 算法以及改进后的 SEaTH 算法对实验区进行面向对象分类，分类结果如图 5 所示。另外，为定量评价两个分类结果，从地面采样点和 Google Earth 高分辨率影像上选取了 190 个验证点，对结果进行了混淆矩阵的计算。两个分类结果的混淆矩阵分别如表 4、表 5 所示。

由分类结果图(图 5)可以看出，虽然原始 SEaTH 算法也能够将大部分地物提取出来，但是在人工表面边缘以及林地边缘都与耕地有所混淆(图 5 中黑色圆圈区域)，而耕地与林地和人工表面的精确区分一直是利用单时相影像进行分类的难点。改进后的 SEaTH 算法在一定程度上改善了这个问题，提高了耕地和林地的分类精度，尤其是耕地的精度提高了 12.26%。对耕地精度提高的原因进行分析，一方面，对两种方法筛选出的分类特征进行分析可以发现，原始 SEaTH 算法中用于区分耕地与其他地物的

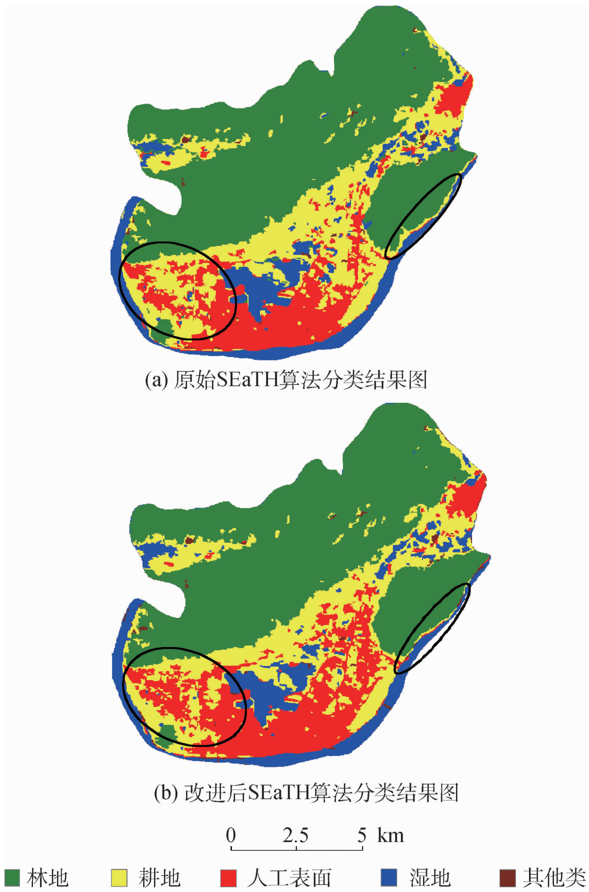


图 5 改进前后 SEaTH 算法分类结果图

表 4 原始 SEaTH 算法分类结果混淆矩阵

样本数据		参考数据						
		林地	耕地	人工表面	湿地	其他	总和	使用精度/%
检测结果	林地	53	1	0	1	1	56	94.64
	耕地	7	28	16	5	0	56	50.00
	人工表面	0	5	50	2	0	57	87.72
	湿地	0	0	0	20	0	20	100
	其他	0	0	0	0	1	1	100
	总和	60	34	66	28	2	190	
	生产精度/%	88.33	82.35	75.76	71.43	50.00		
	总体精度=80.00% Kappa=0.7277							

表 5 改进 SEaTH 算法分类结果混淆矩阵

样本数据		参考数据						
		林地	耕地	人工表面	湿地	其他	总和	使用精度/%
检测结果	林地	53	0	1	0	1	55	96.36
	耕地	7	33	11	2	0	53	62.26
	人工表面	0	1	54	5	0	60	90.00
	湿地	0	0	0	21	0	21	100
	其他	0	0	0	0	1	1	100
	总和	60	34	66	28	2	190	
	生产精度/%	88.33	97.06	81.82	75.00	50.00		
	总体精度=85.26% Kappa=0.7989							

特征主要集中于 TM 影像的第 3、4、5、7 波段以及 DEM,特征种类相对单一,重复性较大;而改进后的 SEaTH 算法筛选出的特征不仅含有原始光谱特征,更包括光谱运算后得到的比值居民地指数特征、纹理特征、亮度特征等其他多元特征,从不同角度对地物进行了区分,起到了很好的互补作用。另一方面,对混淆矩阵进行分析可以看出,第 2 种方法的分类结果中耕地与人工表面以及与湿地的错分误差减少了,分别由原来的 28.57% 和 8.93% 减少到 20.75% 和 3.77%,漏分误差由原来的 17.65% 降低到 2.94%,整体的分类精度得到提高。

然而,与其他类别相比,耕地的分类精度仍然较低,并且较大程度上与人工表面混淆。这主要是因为用于提取耕地的特征主要是从 9 月和 12 月份的影像中得到的,而在此时段肇庆的耕地大部分处于收割期,因此反映出的是裸露地表的特性,很容易与城市边缘正在建设的裸地相混淆,这也是利用单一时相的遥感影像区分耕地与人工表面的困难之处。其他 4 种地物的分类精度都在 90% 以上,这也与验证点的分布有关。总体看来,改进后的 SEaTH 算法分类精度可以达到 85.26%,已经基本满足自动分类的精度要求,并且使林地、耕地和人工表面 3 类地物的分类精度都得到提高,总体精度提高了 5.26%,Kappa 系数提高了 0.0712,说明改进后的算法提取的特征分类效果更好。

4 结论与展望

在 SEaTH 算法的基础上引入了去相关以及类内距离两个指标,提出了一种新的面向对象分类的特征筛选方法,并将此方法运用到了中低分辨率的遥感土地覆盖分类中。通过对比实验可以得到以下结论:

(1)加入去相关分析可以使筛选出的分类特征包含的信息量更充分,并且避免信息冗余。

(2)改进方法筛选出的特征更多元化,除了光谱特征外还加入了纹理特征、最大差分和亮度等特征。

(3)对中低分辨率遥感影像,形状特征对区分地物并不十分有效。

(4)改进方法能够有效提高人工表面、林地和耕地的分类精度,而这三者往往是遥感影像分类的难点,尤其对于云雨较多的地区,往往不易获取质量好的多时相数据,本方法的提出对单时相的遥感影像分类具有重要意义。

另外,利用此种分类方法时对样本的选择有一定要求,即选择的样本要具有代表性,因为选择不同的样本最后筛选出的特征可能完全不同,这直接影响了分类的结果。总体来看,本文提出的自动分类方法可以大大提高土地覆盖分类工作的效率,并且减少了人为主观因素的影响,对大面积、多源遥感数据的土地覆盖分类工作具有重要意义。

然而,这种方法也存在着以下不足:(1)只对一级分类系统较为有效,而对于更为细致的类别(如林地中阔叶林与针叶林)的区分效果还需要进一步验证。(2)本文中得到的耕地的分类精度并不是很高,还需要进一步研究提高其分类精度,以使用于更加深入的研究中。另外,在今后的工作中还应该进一步研究分类顺序对结果的影响。

参考文献(References)

Addink E A, de Jong S M, Davis S A, Dubyanskiy V, Burdelow L A and Leirs H. 2010. The use of high-resolution remote sensing for plague surveillance in Kazakhstan. Remote Sensing of Environment, 114

(3): 674−681 [DOI: 10.1016/j.rse.2009.11.015]

Bruzzone L and Serpico S B. 2000. A technique for feature selection in multiclass problems. International Journal of Remote Sensing, 21(3): 549−563 [DOI: 10.1080/014311600210740]

陈杰, 邓敏, 肖鹏峰, 杨敏华, 梅小明. 2010. 粗糙集高分辨率遥感影像面向对象分类. 遥感学报, 14(6): 1139−1155

冯莉, 李满春, 李飞雪. 2008. 基于遗传算法的遥感图像纹理特征选择. 南京大学学报(自然科学), 44(3): 310−319

高伟. 2010. 基于特征知识库的遥感信息提取技术研究. 北京: 中国地质大学

Gao Y, Marpu P, Niemeyer I, Runfola D M, Giner N M, Hamill T and Pontius R G Jr. 2011. Object-based classification with features extracted by a semi-automatic feature extraction algorithm-SEaTH. Geocarto International, 26(3): 211−226 [DOI: 10.1080/10106049.2011.556754]

韩鹏, 龚健雅, 李志林, 柏延臣, 程亮. 2010. 遥感影像分类中的空间尺度选择方法研究. 遥感学报, 14(3): 507−518

何小晨, 徐守时. 2002. 基于关联规则的特征选择方法. 红外与激光工程, 31(6): 504−507

吉小军, 李世中, 李霆. 2001. 相关分析在特征选择中的应用. 华北工学院测试技术学报, 15(1): 15−18

Koller D and Sahami M. 1996. Toward Optimal Feature Selection // Proceedings of International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann: 284−292

Laliberte A S, Browning D M and Rango A. 2012. A comparison of three feature selection methods for object-based classification of sub-decimeter resolution UltraCam-L imagery. International Journal of Applied Earth Observation and Geoinformation, 15: 70 − 78 [DOI: 10.1016/j.jag.2011.05.011]

Nussbaum S, Niemeyer I and Canty M J. 2006. SEaTH-a new tool for automated feature extraction in the context of object-based image analysis // Proceedings of the 1st International Conference on Object-based Image Analysis. Austria: Salzburg University

Peng H, Long F and Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8): 1226−1238 [DOI: 10.1109/TPAMI.2005.159]

孙即祥. 2002. 现代模式识别. 北京: 高等教育出版社

吴波, 朱勤东, 高海燕, 周小成. 2009. 面向对象影像分类中基于最大化互信息的特征选择. 国土资源遥感, 81(3): 30−34

吴宏安, 蒋建军, 张海龙, 张丽, 周杰. 2006. 比值居民地指数在城镇信息提取中的应用. 南京师大学报(自然科学版), 29(3): 118−121

叶志伟, 郑肇葆, 万幼川, 虞欣. 2007. 基于蚁群优化的特征选择新方法. 武汉大学学报(信息科学版), 32(12): 1127−1130

Zeng Z Y. 2007. A new method of data transformation for satellite images: I. Methodology and transformation equations for TM images. International Journal of Remote Sensing, 28(18): 4095−4123 [DOI: 10.1080/01431160601028912]

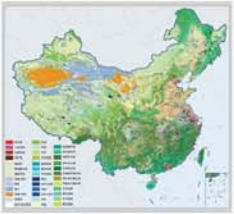
张秀英, 冯学智, 江洪. 2009. 面向对象分类的特征空间优化. 遥感学报, 13(4): 664−669

张振平. 2006. 基于 Bayes 错误率上界最小的特征选择算法的研究. 上海: 同济大学

郑毅. 2010. 基于 SPOT5 全色影像的土地利用信息提取研究. 北京: 中国地质大学

遥感学报

启功先生创刊题名



封面说明

About the Cover

2010年中国土地覆被遥感监测数据集（ChinaCover2010）

The China National Land Cover Data for 2010 (ChinaCover2010)

2010 年中国土地覆被遥感监测数据集（ChinaCover2010）由中国科学院遥感与数字地球研究所联合其他 9 个单位历时两年完成，应用 30 m 空间分辨率的环境星（HJ-1A/1B）数据，利用联合国粮农组织（FAO）的 LCCS 分类工具，构建了适用于中国生态特征的 38 类土地覆被分类系统，采用基于超算平台的数据预处理、面向对象的自动分类、地面调查获得的 10 万个野外样本以及雷达数据辅助分类相结合的方法，数据精度达到 85%。ChinaCover2010 主要基于国产卫星影像，将遥感与生态紧密结合，充足的野外样点以及严格的产品质量控制在最大程度上保证了数据的精度，可为中国生态环境变化评估以及生态系统碳估算提供基础数据支撑。（网址：http://www.chinacover.org.cn）

The China National Land Cover Data for 2010 (ChinaCover2010) has been completed after two years of team effort by the Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences (CAS), together with nine other institutions' participation. The HJ-1A/1B satellite at 30 m resolution is main data source. Based on the landscape features in China, 38 land cover classes have been defined using UN FAO Land Cover Classification System (LCCS). Super computers were used in the data preprocessing. An object-oriented method and a thorough field survey (about 100000 field samples) were used in the land cover classification, with radar imagery as auxiliary data. The overall accuracy of ChinaCover2010 is around 85%. Mainly based on domestic imagery, the products take advantage of various in situ data and strict quality control. ChinaCover2010 is a good dataset for ecological environment change assessment and terrestrial carbon budget studies. (Website: http://www.chinacover.org.cn)

遥感学报

JOURNAL OF REMOTE SENSING

YAOGAN XUEBAO（双月刊　1997年创刊）

第17卷　第4期　2013年7月25日

(Bimonthly, Started in 1997)

Vol.17 No.4 July 25, 2013

主　管	中国科学院	Superintended by	Chinese Academy of Sciences
主　办	中国科学院遥感与数字地球研究所 中国地理学会环境遥感分会	Sponsored by	Institute of Remote Sensing and Digital Earth,CAS The Associate on Environment Remote Sensing of China
主　编	顾行发	Editor-in-Chief	GU Xing-fa
编　辑	《遥感学报》编委会 北京市安外大屯路中国科学院遥感与数字地球研究所 邮编：100101　电话：86-10-64806643 http://www.jors.cn E-mail:jrs@irsa.ac.cn	Edited by	Editorial Board of Journal of Remote Sensing Add: P.O.Box 9718, Beijing 100101, China Tel: 86-10-64806643 http://www.jors.cn E-mail: jrs@irsa.ac.cn
出　版	科学出版社	Published by	Science Press
印刷装订	北京科信印刷有限公司	Printed by	Beijing Kexin Printing Co. Ltd.
总发行	科学出版社 北京东黄城根北街16号 邮政编码：100717 电话：86-10-64017032 E-mail:sales_journal@mail.sciencep.com	Distributed by	Science Press Add: 16 Donghuangchenggen North Street, Beijing 100717, China Tel: 86-10-64017032 E-mail: sales_journal@mail.sciencep.com
国外发行	中国国际图书贸易总公司 北京 399 信箱　邮政编码：100044	Overseas distributed by	China International Book Trading Corporation Add: P.O.Box 399, Beijing 100044, China

中国标准连续出版物号：ISSN 1007-4619 / CN 11-3841/TP　国内邮发代号：82-324　国外发行代号：BM 1002　定价：70. 00元

CODEN YXAUAB

ISSN 1007-4619

国内外公开发行