doi:10.6041/j.issn.1000-1298.2016.01.040

Classification of Land Use in Farming Area Based on Random Forest Algorithm

Ma Yue[1] Jiang Qigang[1] Meng Zhiguo[1,2] Li Yuanhua[1] Wang Dong[3] Liu Huaxin[1]

(1. *College of Geo-exploration Science and Technology*, *Jilin University*, *Changchun* 130026, *China*
2. *Key Laboratory of Planetary Sciences*, *Chinese Academy of Sciences*, *Shanghai* 200030, *China*
3. *College of Forestry*, *Southwest Forestry University*, *Kunming* 650224, *China*)

Abstract: Land use classification plays an important role in adjusting land structure and developing land resources reasonably, especially in the farming area. The objective of this research is to choose an appropriate method to classify land use type in the farming area. A new classification method, Random Forest (RF) classifier, was applied to make land use mapping in agricultural cultivation region with multi-source information, including multi-seasonal spectrum, texture and topographic information. The best classification scheme was chosen to extract land use information, and RF was used to reduce the dimension of characteristics variables. The RF, support vector machine, and maximum likelihood classification were used to map agricultural land use, and the applicability of these three different classification methods was analyzed. The result shows that RF classification of land use classification with multi-source information effects best, the overall accuracy and Kappa coefficient were 85.54% and 0.835 9 respectively. Feature selection method from RF can effectively reduce the data dimension and ensure the accuracy of classification at the same time. Compared with these three classification methods, RF performs the highest overall accuracy of 81.08%, which was respectively 9.46% and 5.27% higher than support vector machine and maximum likelihood classification. It is an effectively scheme that makes land use classification in the farming area using RF classifier with multi-source information. It provides a fast and feasible method for the division of land use types.

Key words: land use classification; farming area; random forest algorithm; multi-source information; feature selection

0 Introduction

Land use classification plays an important role in adjusting land use structure, developing land resources and monitoring land use status[1-2]. Remote sensing, with the fast, macroscopic and synchronous advantages, can provide efficient and rapid technical means in extracting land use information[3-5]. Combined with the remote sensing data and machine learning algorithms has being a research focus in classification field. Many researchers try to develop the performance of different classifiers, such as maximum likelihood classification, neural networks and support vector machine, and obtain good classification results[6-9].

A promising classifier called random forest(RF) has been widely applied in many fields in overseas, however, Chinese researchers have paid little attention to use this method, especially in mapping land use classes of farming area[10-19]. The aim of this paper is to apply RF to make land use mapping in agricultural cultivation region with multi-source information, including multi-seasonal spectrum, texture and topographic information. We first analyzed the influence of classification results by adding many different variables and discussed the importance of variables. RF classifier was applied to reduce the number of input variables by variables importance to improve classification efficiency. Then, RF was compared with support vector machine and maximum likelihood classification in classification performance. Finally, we assessed the availability of classification scheme which applied RF classifier with multi-source information in land use classification of the farming

Received date: 2015-06-16 Accepted date: 2015-08-07
Supported by National Natural Science Foundation of China (Grant No. 41371332), China Geological Survey(1212011220105), and Macao Science and Technology Development Fund(110/2014/A3)
Corresponding author: Jiang Qigang, Professor. E-mail: jiangqigang@jlu.edu.cn. Tel: +86-13180800128.

area. This scheme is of great significance in monitoring land-use status and managing land resources in agricultural cultivation region in the future.

1 Materials

1.1 Study area

The study area is located at 47°0′5″N ~ 47°27′4″N, 123°51′12″E ~ 124°29′47″E, in the southeast of Heilongjiang Province. It includes an agricultural cultivation region nearby Qiqihar and connects to Zhalong wetland which is about 26.7 km away from the south of Qiqihar (Fig. 1). Land use types in study area including water area and water conservancy facilities (lake, river, pond, canal and shoaly land), cultivate land (paddy field and upland field), residential land (habitation), and other lands (alkali-saline land and marsh), there is no woodland and meadow. During classification experiment, we took the water area, water conservancy facilities and cultivate land as major classified objects, the habitation and other land types as secondary objects, besides we did not consider the woodland and meadow.

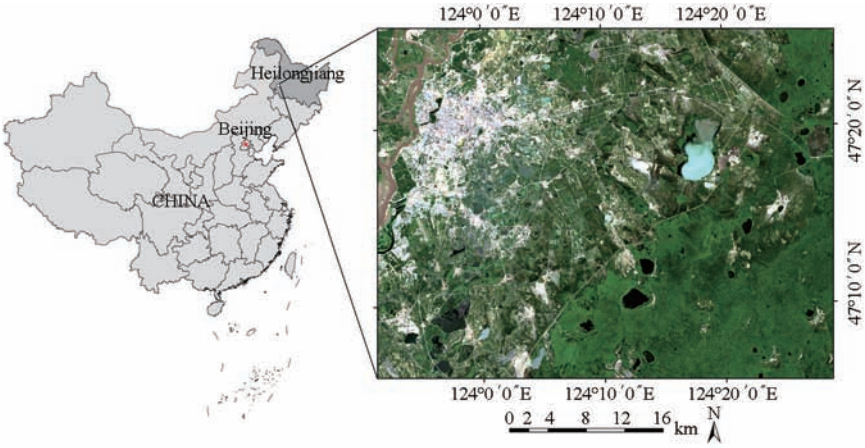

Fig. 1 Study area

1.2 Data sets

Three Landsat8 (OLI) images from different seasons were chosen as the multi-temporal input data. The spring, summer, and autumn images were respectively acquired on April 21st 2013 (LC81200272013111LGN01), July 10th 2013 (LC81200272013191LGN00), and October 30th 2013 (LC81200272013303LGN00). All these multispectral images were quality and cloudless with 30m spatial resolution. The topographic data produced by USGS NASA and METI in June 29th 2009 was ASTER Global DEM. And the 30m DEM has been re-projected to UTM/WGS84.

2 Methods

2.1 Data processing

For spectral data: first, three multispectral Landsat8 images were conducted radiometric calibration and atmospherically corrected using FLAASH module in ENVI 5.1. Then corrected data was subset in order to consistent with the scope of study area. Finally, these multi-spectral images were re-projected to the UTM coordinate system in zone 51 (WGS84 datum) according to the location of study area.

For topographic data: firstly, two GDEM images were mosaicked to a complete elevation image. Secondly, the new mosaic image was cut in accordance with the size of study area. Finally, DEM and spectral data were co-registered together keeping same geographic reference. For using optical data and topographic data in classification at the same time, these two dataset must both resample to 30 m spatial resolution.

2.2 Feature extraction

Spectrum feature variable: we extracted different predictive variables from three seasonal Landsat8 (OLI) images (spring, summer and autumn), including band2 to band7, first two principal component (PCA1 and PCA2), normalized difference vegetation index (NDVI) and modified normalized difference water index (MNDWI)[20].

Texture feature variable: due to a variety of object types with obvious texture features in summer, the first principal component from summer multispectral image was used to calculate eight statistical texture features by gray level co-occurrence matrix (GLCM), including mean, variance, homogeneity, contrast, dissimilarity,

entropy, second moment and correlation[21].

Topographic feature variable: this type of feature variables were derived from GDEM image, which included elevation, slope and aspect information. All of the feature variables used in classification have been shown as follow in Tab. 1.

Tab. 1 Statistics of characteristic parameters

Feature information		Feature variables	Count
Spectrum(OLI)	Spring(SPR)	B,G,R,NIR,SWIR1,SWIR2,PCA1,PCA2,MNDWI,NDVI	10
	Summer(SUM)	B,G,R,NIR,SWIR1,SWIR2,PCA1,PCA2,MNDWI,NDVI	10
	Autumn(AUT)	B,G,R,NIR,SWIR1,SWIR2,PCA1,PCA2,MNDWI,NDVI	10
Texture(TXT)		Mean (mea), Variance (var), Homogeneity (hom), Contrast (con), Dissimilarity (dis), Entropy (ent), Second moment (sm), Correlation (cor)	8
Topography(DEM)		Elevation(eleva), Slope, Aspect	3

2.3 Classification technique

Random forest is a promising machine learning algorithm which consists of many CART decision trees[22]: first, N training data are randomly extracted from original data by bootstrap sampling, and each of the training set is two-thirds of original data. Then, N classification and regression trees (CARTs) which are generated using bootstrap samples can grow into a classification forest. With trees growing, m features in the total M variables ($m \leqslant M$) are randomly sampled and individual trees uses the Gini Index as a measure to select the best variable from m features for node splitting. Finally, each CART produces a predictive result as a vote, and one class is determined by a majority vote among the N CARTs. About one-third samples are not selected to construct CARTs called out-of-bag (OOB) data. OOB data can be used to evaluate internal error of random forest classifier called out-of-bag error.

Random forest algorithm was performed in R-project. There are two input parameters need to be defined in this algorithm, one is the number of trees (N), the other is the number of split variables (m). The experiments showed that OOB error trends to be stable when N is more than 500, and random forest did not appear over-fit phenomena. In this study, we set N to 500, and the square root of the total number of variables ($\sqrt{M}$) was used as the number of split variables (m) at the nodes.

2.4 Importance of the variables and feature selection

Random forest variable importance is calculated by considering the OOB error: first, the out-of-bag error of each decision tree (e_t) is calculated according to the out-of-bag data. Then the value of variable X^j is randomly permuted and out-of-bag error is recalculated (e_t^j). Finally, variable importance of X^j is equal to[23-24]

$$V(X^j) = \frac{1}{N}\sum_{t=1}^{N}(e_t^j - e_t)$$

Due to the change of variable X^j, if the more the out-of-bag error increases and the greater the classification accuracy reduces, it explains that the more important the variable X^j is.

In this paper, we input a large number of feature variables for classification, such as spectral variables, texture variables and topographic variables, but not each of them can play a significantly positive role on the classification. With the advantages of random forest algorithm, these feature variables are selected by the variable importance to reduce the dimension of the input data set.

2.5 Accuracy assessment

In order to assess the accuracy of different classification results, we randomly collected 740 samples to set up an accuracy assessment database using high-resolution image from Google Earth. The confusion matrix and precision index were calculated to verify the classification accuracy.

3 Results and discussion

3.1 Results of classification schemes

3.1.1 Accuracy assessment among different schemes

To choose the best scheme for land use classification, six models, SPR, SUM, AUT, OLI, OLI + TXT and OLI + TXT + DEM, were explored in our research. A confusion matrix was calculated to estimate the overall accuracy, Kappa coefficients,

omission error rates and commission error rates for each classification model. The different results of each model were presented in error figures (Fig. 2).

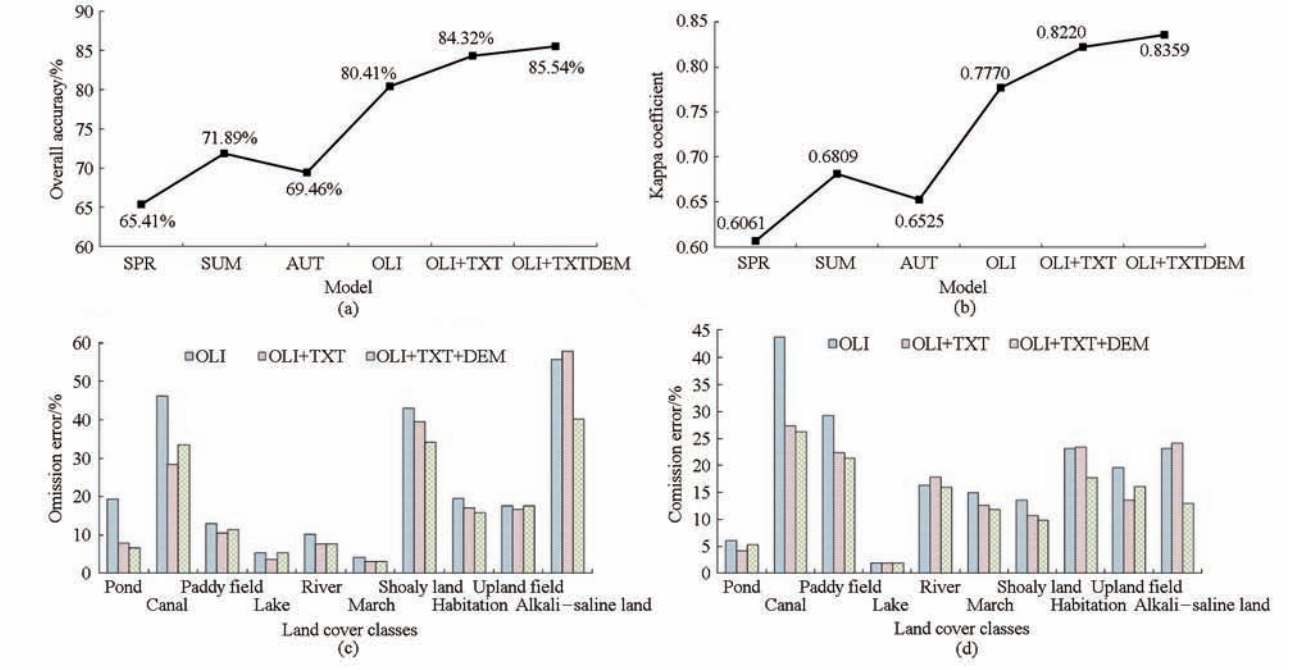

Fig. 2 Comparison of errors

From Figs. 2a and 2b, compared the overall accuracy and Kappa coefficients between six models, the accuracy rate gradually increased with adding different feature variables in models. Adding spectrum feature variables, overall accuracy and Kappa coefficients significantly increased from single-season spectral model (SPR, SUM, AUT) to multi-season spectral model (OLI). The overall accuracy and Kappa coefficients were highest rose by 15% and 0.170 9 respectively. Adding texture and topographic feature variables, the overall accuracy of OLI, OLI + TXT and OLI + TXT + DEM models were 80.41%, 84.32% and 85.54% respectively, successively rose by 3.91% and 1.22%. According to the rising range of accuracy rates, texture variables had a greater influence on classification accuracy than topographic variables.

From Figs. 2c and 2d, it shows the commission error rates and omission error rates from three classification models (OLI, OLI + TXT and OLI + TXT + DEM), of which the overall accuracy and Kappa coefficients were higher than others among the six predictive models. In the OLI + TXT model with adding texture variables, both the commission error rates and omission error rates of pond, canal, paddy field, shoaly land and marsh were dropped, especially of the pond, canals and paddy field. The OLI + TXT + DEM model adding topographic variables, except for the pond and paddy field, commission error rates of classification objects were all lower than OLI + TXT model.

3.1.2 Importance of the variables

As shown in Fig. 3, the first twenty-five variables from OLI, OLI + TXT and OLI + TXT + DEM were respectively listed according to the OOB error. In the figure, spectrum variables were described as 'Band name_season'. The ndvi, mndwi, b2, b4, b5, b6, b7 and pca2 respectively represented normalized difference vegetation index, modified normalized difference water index, blue band, red band, near-infrared band, shortwave infrared band1, shortwave infrared band2 and second principal component, which were extracted from Landsat8 (OLI). The spr, sum and aut respectively represented the season of spring, summer and autumn. The con, mea, dis, ent, var and hom respectively represented contrast, mean, dissimilarity, entropy, variance and homogeneity, which were calculated by gray level co-occurrence matrix (GLCM). The eleva represented elevation from DEM image.

From Fig. 3a, in OLI model, the three-season variables all contributed to improve the classification accuracy. The variables of ndvi_sum, mndwi_sum, b7_sum, b4_sum and b6_sum effected more important in classification, and ranked in the top five. As for

OLI + TXT model, in Fig. 3b, con (contrast) was the most important variable in OLI + TXT model, followed by mea (mean), dis (dissimilarity), ent (entropy) and var (variance). In the OLI + TXT + DEM model, as Fig. 3c shows, among the topographic variables, eleva (elevation) played a major role in improving the classification accuracy.

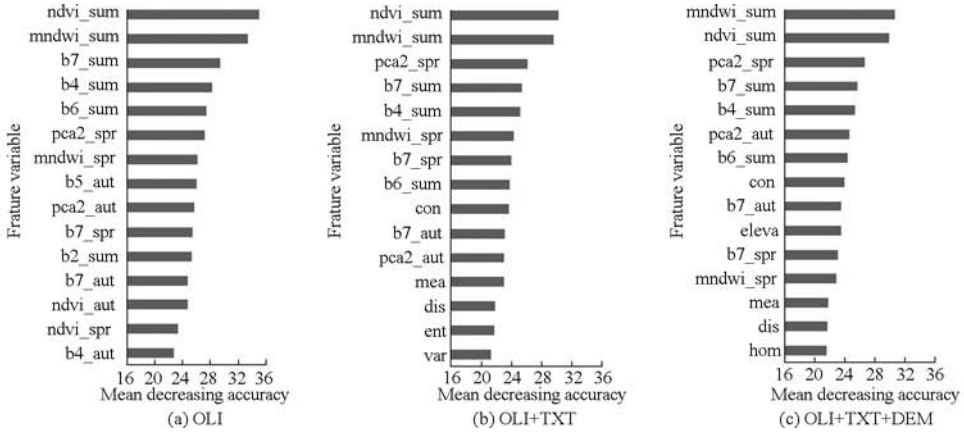

Fig. 3 Variable importance

3.2 Discussion of classification schemes

We analyzed and estimated six classification models (SPR, SUM, AUT, OLI, OLI + TXT and OLI + TXT + DEM) based on the four evaluation indexes, including overall accuracy, Kappa coefficients, omission error rates and commission error rates. This research found that, the model with multi-source information (OLI + TXT + DEM), including multi-seasonal spectrum, texture and topographic variables, produced the highest classification accuracy, it could effectively distinguish the agricultural land and non-agricultural land, and extract the land use information more accurately in the farming area.

Moreover, in order to confirm how did feature variables added in models influence on the classification accuracy, we compared the importance of feature variables calculated by random forest algorithm and the characteristics of land use types. The results showed that among the classification types of land use in the study area, the paddy field, upland field and shoaly land had more information about vegetation and water. And the variables, such as MNDWI, NDVI, red and near infrared band which were related to the vegetation-water information, were more important in classification than other feature variables. The result of feature importance derived from random forest algorithm was consistent well with the actual situation. As for the water conservancy facilities, such as pond, canal and paddy field, these land cover classes had obvious texture characteristics. And land cover types, such as river, lake, upland field and marsh, had significant differences on the texture homogeneity. The location of farming area was less affected by the topographic relief, and the importance of texture variables calculated by random forest was higher than the topographic variables. The result presented that random forest feature importance was consistent well with the actual situation on texture and topographic characteristic. In general, random forest could well use the differences between samples, select the most relative information of land cover types, and then classify by the best variables. The importance of variables from random forest algorithm could be reliably used to extract the feature variables.

4 Comparison of classifiers

4.1 Feature selection

In general, model with multi-source information (OLI + TXT + DEM) produced the highest classification accuracy, but it added too many feature variables in the classification. After analysis of the variable importance from random forest algorithm, among the 41 variables added in the OLI + TXT + DEM model, we selected the first ten feature variables which well contributed to the classification accuracy and contained much different feature information. These ten variables were mndwi_sum, ndvi_sum, pca2_spr, b7_sum, b4_sum, pca2_aut, b6_sum, con, b7_aut and eleva. In the research, feature information was selected by the importance of variables, random forest

classifier could automatically remove the less important variables to reduce the dimension of variables and time cost, and improve the classification efficiency.

4.2 Accuracy assessment among different classifiers

To compare the applicability of different classifiers on extracting land use types, we applied three classifiers with selected feature variables, the classifiers included random forest, maximum likelihood classification and support vector machine, which were marked as FS _ RF, FS _ MLC and FS _ SVM. A confusion matrix was calculated according to the classification results to estimate the overall accuracy and Kappa coefficients. In order to verify the differences of classification accuracy between features selected and non-selected, we compared the results of classification with feature selection and without feature selection (Tab. 2).

From Tab. 2, overall accuracy and Kappa coefficients of FS _ RF model were 81.08% and 0.785 2, respectively, little different from the classification result without feature selection, and the overall rates could also reach more than 80%. But with feature selection, the dimensionality of variables was greatly reduced from 41 variables to 10 variables. Through the above analysis, a classification scheme of random forest algorithm with feature selection could remain the most important information of classification objects. This scheme could reduce the dimensionality of input datasets to shorten operation time and improve classification efficiency, but it could also produce higher accuracy of classification.

Among the three classification models, FS _ RF model produced the highest overall accuracy of 81.08% and Kappa coefficients of 0.785 2. The research found that random forest algorithm could process high-dimensional data in parallel to shorten operation time and improve the efficiency of classification. The results showed that random forest could effectively distinguish agricultural land and non-agricultural land, and perform good applicability of extracting the land cover information more accurately in farming area.

Tab. 2 Comparison of different classification accuracies

Index	OLI_RF	OLI + TXT_RF	OLI + TXT + DEM_RF	FS_RF	FS_SVM	FS_MLC
Overall accuracy/%	80.41	84.32	85.54	81.08	71.62	75.81
Kappa coefficient	0.777 0	0.822 0	0.835 9	0.785 2	0.727 6	0.678 3

4.3 Comparison of classification results

In order to reveal the results of land use classification with variables selection by different classifiers in the farming area, we analyzed different results of 4 classification models, OLI + TXT + DEM_ RF, FS_RF, FS_SVM and FS_MLC, and then land use mapping was present as Fig. 4 showed.

From Figs. 4a and 4b, after selecting variables, random forest algorithm could also clearly distinguish every land cover type in the farming area. Compared with the scheme of classification with high-dimensional variables, RF with variable selection could more efficiently and rapidly extract the information of land use cover. From Figs. 4b, 4c and 4d, among the three classifiers, support vector machine and maximum likelihood classification algorithm made some obvious mistakes during classification. For example, in Fig. 4c, many ponds were classified as lakes by mistake and in Fig. 4d, the cultivated land was seriously mistook as canals in the classification result. In a word, it was a reasonable classification scheme of using random forest algorithm with selected variables, which truly realized to extract information efficiently.

5 Conclusions

Land use cover could achieve to classify in the farming area based on random forest algorithm with multi-source information, including multi-seasonal spectrum, texture and topographic information. Compared with the traditional classifier, random forest classifier could effectively apply the differences between training samples, and select the best variable to classify. The ability of processing high-dimensional data in parallel significantly improved the efficiency of land use classification in farming area. Random forest variable importance could be used as a measure to select features, and effectively reduced the dimensionality of input data. Combined with the comprehensive information with selection and random forest algorithm, it could truly realize a good balance

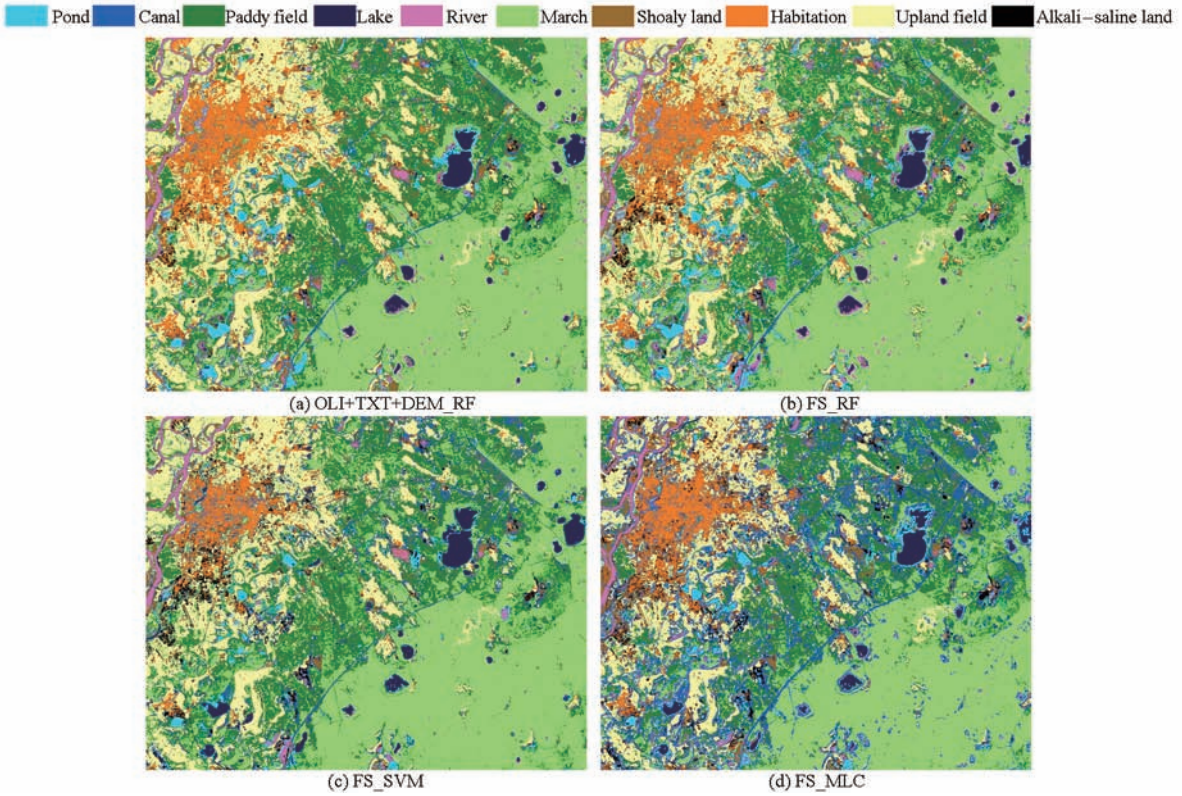

Fig. 4 Comparison of different classification results

between classification efficiency, accuracy, and applicability. It provided a reliable method and advantageous basis to extract the information of land use in a large area of agricultural cultivation region in the future.

References

[1] Lin Nan, Jiang Qigang, Yang Jiajia, et al. Classifications of agricultural land use based on high-spatial ZY1 - 02C remote sensing images [J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1):278 - 284. (in Chinese)

[2] Li Jiangang. Coordination of research in different of land use classification criteria [D]. Beijing: China University of Geosciences, 2012. (in Chinese)

[3] Wei Jiwei. The study of land use classification based on remote sensing image [D]. Changchun: Northeast Normal University, 2012. (in Chinese)

[4] Sun Danfeng, Yang Jihong, Liu Shunxi. Application of high-spatial IKONS remote sensing images in land use classification and change monitoring[J]. Transactions of the CSAE, 2002, 18(2):160 - 164. (in Chinese)

[5] Mai Kaile, Zhang Wenhui. Object-oriented classification approach for remote sensing imagery information extraction in loess hilly-gully region [J]. Transactions of the Chinese Society for Agricultural Machinery, 2011, 42(4):153 - 158. (in Chinese)

[6] Zhao Jianhua. Semi-supervised classification algorithm based on SOM neural network [J]. Journal of Xihua University: Natural Science, 2015, 34(1):36 - 51. (in Chinese)

[7] Meng Zhiguo. The analysis of the application of the BP neural network in the land use classification [D]. Changchun: Jilin University, 2006. (in Chinese)

[8] Zhao Chunhui, Qiao Lei. Classification of hyperspectral remote sensing image using improved LS-SVM [J]. Applied Science and Technology, 2008, 35(1): 44 - 52. (in Chinese)

[9] Fu Wenjie, Hong Jinyi, Lin Mingsen. A method of land use classification from remote sensing image based on support vector machines and spectral similarity scale [J]. Remote Sensing Technology and Application, 2006, 21(1): 25 - 30. (in Chinese)

[10] Jennifer M C, Joseph F K, Alisa L G. Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in northern Minnesota [J]. Remote Sensing, 2013, 5(7):3212 - 3238.

[11] Rodriguez-Galiano V F, Abarca-Hernandez F, Ghimire B, et al. Incorporating spatial variability measures in land-cover classification using random forest [J]. Procedia Environment Sciences, 2011, 3:44 - 49.

[12] Rodriguez-Galiano V F, Chica-Olmo M, Abarca-Hernandez F, et al. Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture [J]. Remote Sensing of

Environment, 2012, 121:93 – 107.

[13] Mellor A, Haywood A, Stone C, et al. The performance of random forests in an operational setting for large area sclerophyll forest classification [J]. Remote Sensing, 2013,5(6):2838 – 2856.

[14] Hayes M M, Miller S N, Murphy M A. High-resolution landcover classification using random forest[J]. Remote Sensing Letters, 2014, 5(2):112 – 121.

[15] Wang Dong, Yue Cairong, Tian Chuanzhao, et al. Classification of TM remote sensing image based on random forests of Dayao Count[J]. Forest Inventory and Planning, 2014, 39(2):1 – 5. (in Chinese)

[16] Li Xinhai. Using random forest for classification and regression [J]. Chinese Journal of Applied Entomology, 2013, 50(4):1190 – 1197. (in Chinese)

[17] Liu Yi, Du Peijun, Zheng Hui, et al. Classification of China small satellite remote sensing image based on random forest [J]. Science of Surveying and Mapping, 37(4): 194 – 196. (in Chinese)

[18] Lei Zhen. Random forest and its application in remote sensing[D]. Shanghai: Shanghai Jiao Tong University, 2012. (in Chinese)

[19] Huang Yan, Zha Weixiong. Comparison on classification performance between random forests and support vector machine[J]. Software, 2012, 33(6):107 – 110. (in Chinese)

[20] Xu Hanqiu. A study on information extraction of water body with the modified normalized difference water index (MNDWI) [J]. Journal of Remote Sensing, 2005, 9(5):589 – 595. (in Chinese)

[21] Wang Shuzhi, Zhang Jianhua, Feng Quan. Defect detection of muskmelon based on texture features and color features [J]. Transactions of the Chinese Society for Agricultural Machinery, 2011, 42(3):175 – 179. (in Chinese)

[22] Breiman L. Random forest [J]. Machine Learning, 2001, 45(1):5 – 32.

[23] Zhu Zhe, Curtis E W, John R, et al. Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat and SAR data [J]. Remote Sensing of Environment, 2012, 117:72 – 82.

[24] Beijma S V, Comber A, Lamb A. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data [J]. Remote Sensing of Environment, 2014, 149:118 – 129.

doi:10.6041/j.issn.1000-1298.2016.01.040

基于随机森林算法的农耕区土地利用分类研究

马　玥[1]　姜琦刚[1]　孟治国[1,2]　李远华[1]　王　栋[3]　刘骅欣[1]

(1. 吉林大学地球探测科学与技术学院, 长春 130026; 2. 中国科学院行星科学重点实验室, 上海 200030;
3. 西南林业大学林学院, 昆明 650224)

摘要: 基于随机森林算法,采用多季节、多时相光谱信息、纹理信息和地形信息进行分类研究,选出最佳分类方案对农耕区土地利用信息进行提取,并利用随机森林算法对所有特征变量进行降维,将降维后的变量分别用于随机森林、支持向量机和最大似然分类法,分析不同分类方法对农耕区土地利用类型提取的适用性。研究结果表明:基于随机森林算法的多源信息综合分类方案进行土地利用信息提取效果最佳,总体精度 85.54%,Kappa 系数 0.835 9;利用随机森林算法进行特征选择可以在有效降低数据维度的同时保证分类精度;3 种分类方法中,随机森林算法总体分类精度 81.08%,分别较支持向量机和最大似然法高 9.46% 和 5.27%。利用随机森林分类法结合多源信息能够有效对农耕区土地利用类型进行分类,为土地类型的划分提供了快捷可行的方法。

关键词: 土地利用分类; 农耕区; 随机森林算法; 多源信息; 特征选择

中图分类号: TP79; F301.24　　**文献标识码**: A　　**文章编号**: 1000-1298(2016)01-0297-07

Classification of Land Use in Farming Area Based on Random Forest Algorithm

Ma Yue[1]　Jiang Qigang[1]　Meng Zhiguo[1,2]　Li Yuanhua[1]　Wang Dong[3]　Liu Huaxin[1]

(1. *College of Geo-exploration Science and Technology*, *Jilin University*, *Changchun* 130026, *China*
2. *Key Laboratory of Planetary Sciences*, *Chinese Academy of Sciences*, *Shanghai* 200030, *China*
3. *College of Forestry*, *Southwest Forestry University*, *Kunming* 650224, *China*)

Abstract: Land use classification plays an important role in adjusting land structure and developing land resources reasonably, especially in the farming area. The objective of this research is to choose an appropriate method to classify land use type in the farming area. A new classification method, random forest (RF) classifier, was applied to make land use mapping in agricultural cultivation region with multi-source information, including multi-seasonal spectrum, texture and topographic information. The best classification scheme was chosen to extract land use information, and RF algorithm was used to reduce the dimension of characteristics variables. The RF algorithm, support vector machine, and maximum likelihood classification were used to map agricultural land use, and the applicability of these three different classification methods was analyzed. The result shows that RF classification of land use classification with multi-source information effects best, the overall accuracy and Kappa coefficient are 85.54% and 0.835 9 respectively. Feature selection method from RF algorithm can effectively reduce the data dimension and ensure the accuracy of classification at the same time. Compared with these three classification methods, RF algorithm performs the highest overall accuracy of 81.08%, which is respectively 9.46% and 5.27% higher than support vector machine and maximum likelihood classification. It is an effective scheme that makes land use classification in the farming area using RF classifier with multi-source information. It provides a fast and feasible method for the division of land use types.

Key words: land use classification; farming area; random forest algorithm; multi-source information; feature selection

收稿日期: 2015-06-16　修回日期: 2015-08-07

基金项目: 国家自然科学基金项目(41371332)、中国地质调查局项目(1212011220105)和澳门科技发展基金项目(110/2014/A3)

作者简介: 马玥(1990—),女,博士生,主要从事遥感地学和环境遥感研究,E-mail: 1191394162@qq.com

通信作者: 姜琦刚(1964—),男,教授,博士生导师,主要从事 GIS 与遥感地学环境研究,E-mail: jiangqigang@jlu.edu.cn

引言

土地利用分类研究在调整土地利用结构、合理开发土地资源、动态监测土地利用状况等环节起着重要作用[1-2]。遥感技术快速、宏观、同步监测等特点,为提取土地利用信息提供高效快捷的技术手段[3-5]。结合遥感数据和机器学习算法进行土地利用分类一直是国内外学者的研究热点,如最大似然、神经网络、支持向量机等方法都被广泛应用,诸多学者对各分类法的不足之处做了完善与改进,得到较好的分类效果[6-9]。

随机森林(Random forest)算法是一种新型高效的组合分类法,其优越的性能在国外诸多领域得到广泛应用,相比而言,国内学者的研究与应用较少,在分类研究中,采用该方法对农业耕种区的土地利用信息提取进行详细探讨的报道较少[10-19]。本文基于随机森林算法,综合多季节、多时相光谱信息、纹理信息和地形信息对农耕区土地利用类型进行分类。分析不同特征信息的加入对土地利用分类结果的影响和各个特征信息在分类过程中的重要程度,并根据随机森林特征变量重要性对高维变量降维,最后将随机森林与最大似然和支持向量机分类法进行比较,评估基于随机森林算法的多源信息综合分类方案在农耕区土地利用分类中的实用性,为监测农耕区土地利用状况、规划管理土地资源提供依据。

1 材料

1.1 研究区概况

研究区位于中国黑龙江省西南部,地理坐标为47°0′5″~47°27′4″N、123°51′12″~124°29′47″E。研究区包括齐齐哈尔市周边的农业耕种区,并与距齐齐哈尔市区南郊 26.7 km 处的扎龙湿地部分地带相连(图 1)。该区域的土地利用类型包括水域及农用水利设施用地(湖泊、河流、库塘、水渠、滩地)、耕地(水田、旱田)、住宅用地(城镇用地)和其他土地(盐碱地、沼泽地)、无林地与草地。本文对研究区进行分类研究时,将水域、农用水利设施用地、耕地作为农业耕种区的主要分类对象,住宅用地、其他土地类型为次要分类对象辅助分析研究,不涉及林地与草地信息的提取。

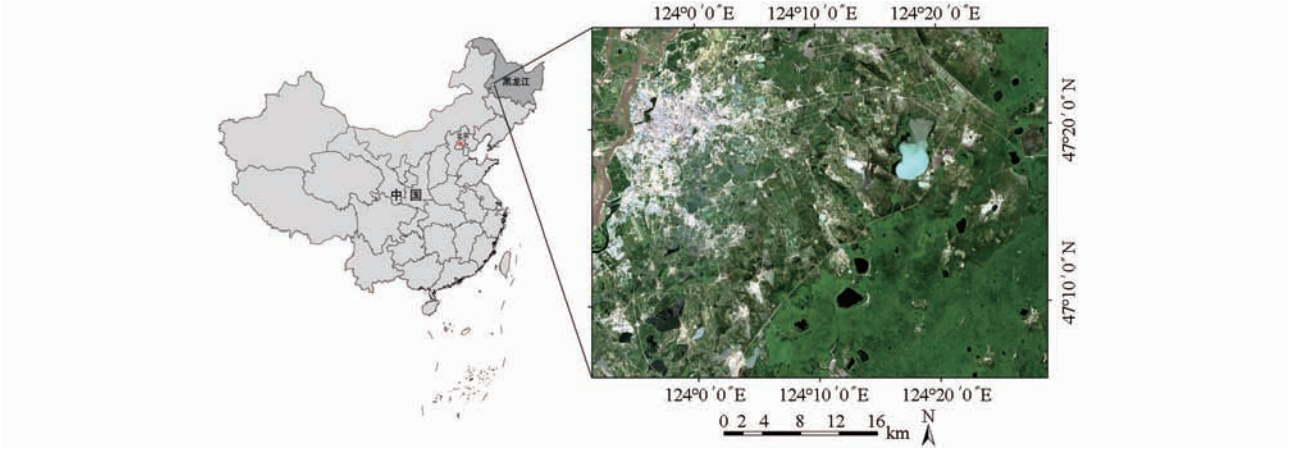

图 1 研究区范围

Fig. 1 Study area

1.2 数据

采用的光谱数据是美国 Landsat8(OLI)卫星于 2013 年 4 月 21 日获取的春季影像(LC81200272013111LGN01),2013 年 7 月 10 日获取的夏季影像(LC81200272013191LGN00)和 2013 年 10 月 30 日获取的秋季影像(LC81200272013303LGN00),空间分辨率 30 m,3 幅影像数据均质量良好,清晰无云;采用的地形数据是美国航空航天局(NASA)和日本经济产业省(METI)于 2009 年 6 月 29 日联合发布的全球数字高程数据产品 ASTER Global DEM。GDEM 数据空间分辨率 30 m,投影为 UTM/WGS84。

2 方法

2.1 数据预处理

多光谱数据:首先利用 ENVI 5.1 中辐射定标功能和 FLAASH 大气校正模型对研究区 Landsat8 影像进行辐射定标和大气校正;然后将校正后影像按照研究区范围进行裁剪;最后根据研究区所在地理位置定义投影参数,本文应用的投影参数为 UTM(zone51)/WGS84。

地形数据:首先将 2 景 GDEM 数据镶嵌成一幅完整的高程影像;其次将 DEM 数据按照研究区范围进行裁剪;最后与多光谱数据统一投影参数。

由于多光谱数据和地形数据要作为不同的变量

参与分类，因此 2 种影像的空间分辨率必须保持一致，均为 30 m。

2.2 特征提取

光谱特征变量：分别提取春、夏、秋 3 个季节 Landsat8（OLI）影像 2～7 波段数据，主成分变换后的前 2 个主成分变量（PCA1、PCA2），归一化植被指数（NDVI），改进的归一化差异水体指数（MNDWI）[20]。

纹理特征变量：夏季（2013 年 7 月 10 日）是研究区内地物种类最全、纹理特征最明显的季节，对夏季多波段遥感影像提取其第 1 主成分，利用灰度共生矩阵（GLCM）统计第 1 主成分的纹理特征，包括均值、方差、同质性、对比度、非相似性、熵、二阶矩、相关性共 8 个参数[21]。

地形特征变量：利用 GDEM 影像提取地形信息，包括高程、坡度和坡向。所有特征提取参数如表 1 所示。

表 1　特征参数统计

Tab. 1　Statistic of characteristic parameters

特征信息		特征参数	个数
光谱信息（OLI）	春季（SPR）	B，G，R，NIR，SWIR1，SWIR2，PCA1，PCA2，MNDWI，NDVI	10
	夏季（SUM）	B，G，R，NIR，SWIR1，SWIR2，PCA1，PCA2，MNDWI，NDVI	10
	秋季（AUT）	B，G，R，NIR，SWIR1，SWIR2，PCA1，PCA2，MNDWI，NDVI	10
纹理信息（TXT）		均值（mea），方差（var），同质性（hom），对比度（con），非相似性（dis），熵（ent），二阶矩（sm），相关性（cor）	8
地形信息（DEM）		高程（eleva），坡度（slope），坡向（aspect）	3

2.3 分类方法

随机森林算法是由多棵 CART 决策树组合构成的新型机器学习算法[22]：首先，采用 bootstrap 抽样技术从原始数据集中抽取 N 个训练集，每个训练集的大小约为原始数据集的 2/3；然后，为每个训练集分别建立分类回归树，产生由 N 棵 CART 决策树组成的森林，在每棵树生长过程中，从全部 M 个特征变量中随机抽选 m 个（$m \leqslant M$），在这 m 个属性中根据 Gini 系数最小原则选出最优属性进行内部节点分支；最后，集合 N 棵决策树的预测结果，采用投票的方式决定新样本的类别；每次抽样约有 1/3 的数据未被抽中，利用这部分袋外数据（Out-of-bag）进行内部误差估计，产生 OOB 误差。

随机森林算法通过 R 语言软件平台实现，在该算法中需要定义 2 个参数：生长树的数目 N 和节点分裂时输入的特征变量个数 m。本文通过实验，当 $N \geqslant 500$ 时，各分类情况的 OOB 误差趋于稳定，随机森林不会出现过拟合现象。本文设置 $N = 500$，节点分裂时输入的特征变量数 $m = \sqrt{M}$ 进行分类。

2.4 特征变量重要性和特征选择

随机森林算法利用 OOB 误差计算特征变量重要性：首先根据袋外数据计算随机森林中每个决策树的袋外误差 e_t；然后随机改变袋外数据第 j 个特征变量 X^j 的值，并计算新的袋外误差 e_t^j；最后变量 X^j 的重要性 $V(X^j)$ 表示为[23-24]

$$V(X^j) = \frac{1}{N}\sum_{t=1}^{N}(e_t^j - e_t)$$

X^j 变量的变化引起的袋外误差增加越大，精度减少的越多，说明该变量越重要。

本文利用光谱、纹理、地形等多种特征变量，但不是每种特征变量都会对分类精度产生显著的作用，根据随机森林算法的特点，提取特征变量的重要性信息，并根据其重要程度，进行特征变量的选择，对高维数据进行降维。

2.5 精度评价

利用 Google Earth 高分辨率同时相影像，随机选取 740 个验证点建立精度评价数据库，对不同分类结果进行精度验证，计算混淆矩阵及相关精度指标。

3 分类方案结果与讨论

3.1 分类方案结果

3.1.1 精度比较

为了选择最佳分类方案对研究区土地利用类型进行分类，实验分为 6 个模型，即 SPR、SUM、AUT、OLI、OLI + TXT、OLI + TXT + DEM，将不同模型的分类结果计算混淆矩阵，得到分类后的总体分类精度、Kappa 系数、各个地物类型的错分误差和漏分误差，并通过误差对比图（图 2）表现各模型分类结果间的差异。

由图 2a、2b，比较 6 个实验模型的总体精度和 Kappa 系数：随着不同类型特征变量的加入，总体精度和 Kappa 系数呈逐步上升趋势。对于加入光谱特征变量，由单季节光谱模型（SPR、SUM、AUT）到多季节光谱模型（OLI），总体精度和 Kappa 系数上升幅度显著，分别最高上升了 15% 和 0.170 9。对于依次加入纹理特征变量和地形特征变量，OLI、OLI + TXT、OLI + TXT + DEM 模型的总体分类精度分别为

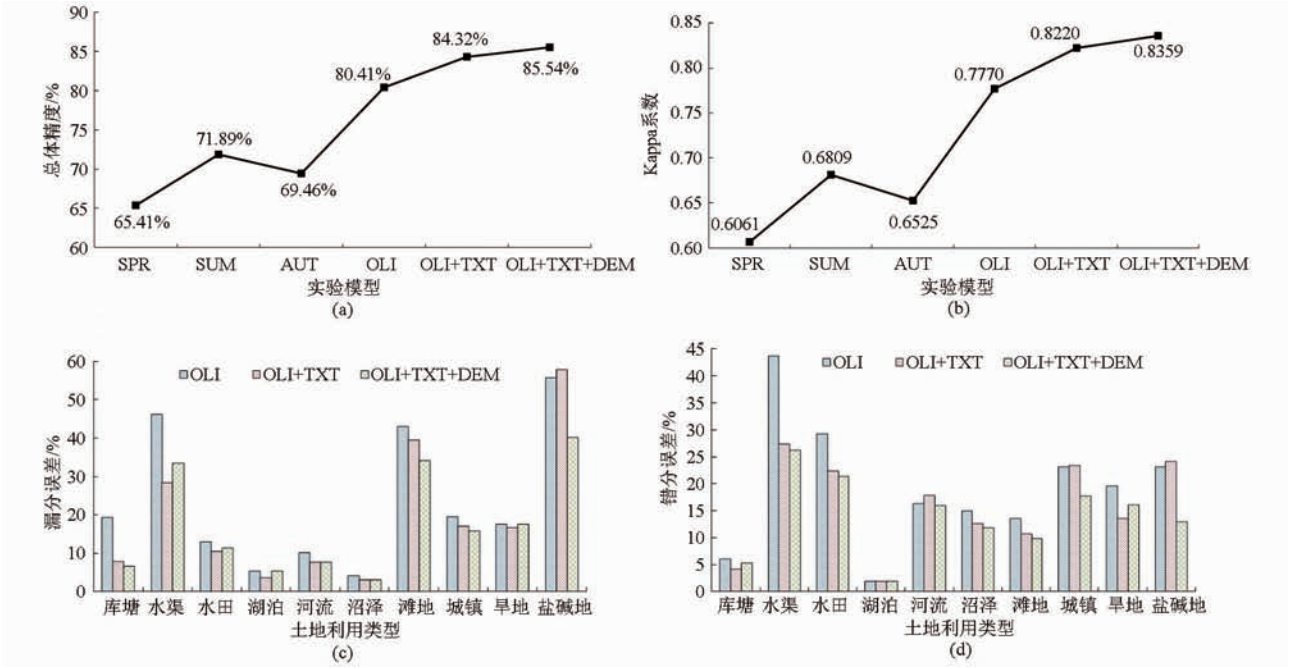

图 2 误差对比

Fig. 2 Comparison of errors

80.41%、84.32%、85.54%,依次上升了 3.91% 和 1.22%,从上升的幅度来看,纹理变量对分类精度的影响大于地形变量。

由图 2c、2d,比较 6 个模型中总体精度和 Kappa 系数较高的 3 个分类模型(OLI、OLI + TXT、OLI + TXT + DEM)下各类地物的漏分误差和错分误差:引入纹理信息的 OLI + TXT 模型,使库塘、水渠、水田、滩地、沼泽的漏分误差和错分误差降低,尤其对于库塘、水渠、水田,纹理特征变量对精度的影响较大,误差下降显著。引入地形信息的 OLI + TXT + DEM 模型,除库塘和旱地外,其他地物类型的错分现象较 OLI + TXT 模型均有所减少。

3.1.2 特征变量重要性

如图 3 所示,根据 OOB 误差分别列出了 OLI、OLI + TXT、OLI + TXT + DEM 模型的前 15 个变量,其中光谱特征变量用"特征波段名称_季节缩写"的形式表述,ndvi、mndwi、b2、b4、b5、b6、b7 和 pca2 分别代表由 Landsat8 遥感影像得到的归一化植被指数、改进的归一化差异水体指数、蓝光波段、红光波段、近红外波段、短波红外 1、短波红外 2 和第 2 主成分变量,季节缩写 spr、sum 和 aut 分别代表春季、夏季和秋季;纹理特征变量 con、mea、dis、ent、var 和 hom 分别代表纹理统计量中的对比度、均值、非相似性、熵、方差和同质性;地形特征变量 eleva 代表高程。

对于 OLI 模型,由图 3a 知,3 个季节变量对分类效果均有贡献,其中 ndvi_sum、mndwi_sum、b7_sum、b4_sum、b6_sum 特征变量的重要程度较大,重要性排在前 5 位。对于 OLI + TXT 模型,由图 3b 知,在参与分类的 8 个纹理参数中,con(对比度)的重要程度最大,其次是 mea(均值)、dis(非相似性)、

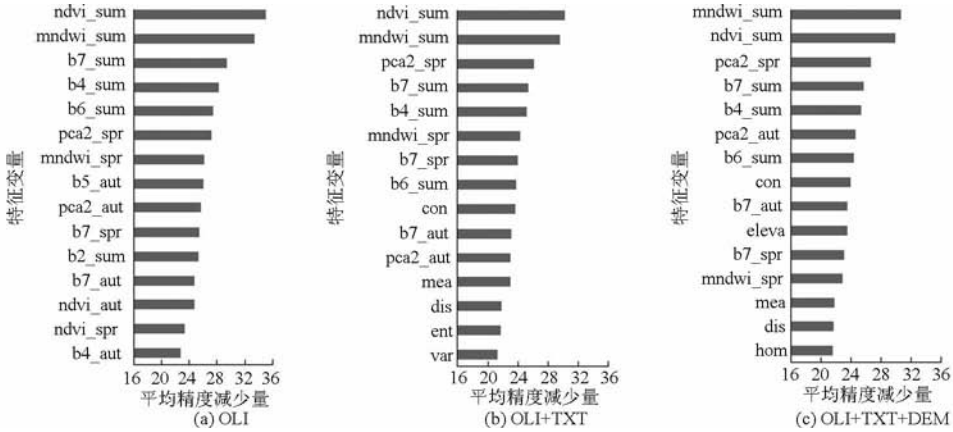

图 3 变量重要性

Fig. 3 Variable importance

ent(熵)和 var(方差)。对于 OLI + TXT + DEM 模型,由图 3c 知,参与分类的地形变量中, eleva(高程)占主要作用,其余变量的重要程度较小。

3.2　讨论

根据总体分类精度、Kappa 系数、各个地物类型的错分误差和漏分误差 4 个评价指标,通过将分类实验的 6 个模型(SPR、SUM、AUT、OLI、OLI + TXT、OLI + TXT + DEM)综合分析与评价发现:在 6 个实验方案中,利用多季节光谱变量、纹理变量和地形变量的多源信息综合模型(OLI + TXT + DEM)分类效果最佳,能够有效区分研究区的农耕用地与非农耕用地,并且对于农耕区内的土地利用类型可以较高精度地进行信息提取。

同时,为了明确各分类变量对分类结果的影响,利用随机森林算法计算各特征变量的重要性,并与研究区土地利用类型特征进行对比分析发现:实验选择的农耕区土地利用类型中,水田、旱地、滩地等地物类型,植被、水体信息较为丰富,与其有关的 MNDWI、NDVI 和红光-近红外波段变量的重要性程度较高,与实际情况基本吻合;库塘、水渠、水田等农用水利设施用地纹理沟纹较深,河流、湖泊、旱地、沼泽等纹理粗细和均匀程度有显著差异,文中农业耕种区所在地理位置地形起伏较小,随机森林算法计算的纹理变量的重要性程度高于地形变量,与实际地物在纹理和地形特征上的差别情况基本吻合。通过分析可知,随机森林算法在分类时可以很好地利用样本间的差异特征,选择与土地利用类型密切相关的信息,根据最优变量进行分类,计算出的变量重要性程度可以作为提取特征变量的依据。

4　分类方法比较

4.1　特征选择

由以上分析可知,多源信息综合的 OLI + TXT + DEM 模型分类精度最高,但应用的特征参数较多。经过随机森林变量重要性分析后,对 OLI + TXT + DEM 模型的 41 个参数进行特征选择,选取对分类贡献最大的前 10 个参数,其中包含了不同的特征信息,分别为 mndwi_sum、ndvi_sum、pca2_spr、b7_sum、b4_sum、pca2_aut、b6_sum、con、b7_aut 和 eleva。通过变量重要性对特征信息进行选择,将重要性小的变量剔除,以降低变量维度,减少模型运算时间,提高效率。

4.2　精度比较

利用经过特征选择后的变量,采用随机森林、最大似然和支持向量机分类法进行分类,分别标记为 FS_RF、FS_MLC 和 FS_SVM,根据分类结果计算混淆矩阵得到总体精度和 Kappa 系数,比较各分类方法对农耕区土地利用类型提取的适用性,为了验证特征选择前后分类精度的差异,将特征选择前后各方案的分类结果进行比较(表 2)。

表 2　不同分类方法精度对比

Tab. 2　Comparison of different classification accuracies

指标	OLI_RF	OLI + TXT_RF	OLI + TXT + DEM_RF	FS_RF	FS_SVM	FS_MLC
总体精度/%	80.41	84.32	85.54	81.08	71.62	75.81
Kappa 系数	0.777 0	0.822 0	0.835 9	0.785 2	0.727 6	0.678 3

由表 2 可知,FS_RF 分类的总体精度和 Kappa 系数分别为 81.08% 和 0.785 2,与特征选择前各分类方案相差不大,总体精度仍可达到 80% 以上,但特征选择后,分类参数的维度大幅度降低,从最高的 41 维减少到 10 维。以上分析说明,基于随机森林算法的特征选择方法,能够保留地物最重要的特征信息,在降低数据维度的同时,分类精度仍能保持较高水平,进而缩短模型运算时间,提高分类效率。

FS_RF、FS_SVM、FS_MLC 3 种分类方法下,随机森林算法(FS_RF)的总体精度和 Kappa 系数最高,分别为 81.08% 和 0.785 2,实验发现,随机森林分类法能够较好地并行处理高维数据,运行时间短,速度快。以上结果显示,随机森林算法较支持向量机和最大似然分类法,更能有效区分农耕用地与非农耕用地,且对于农耕区各土地利用类型提取精度较高,具有很好的适用性。

4.3　分类结果对比

为了分别验证特征选择后的分类效果和不同分类方法对农耕区土地利用分类的影响,选择 OLI + TXT + DEM_RF、FS_RF、FS_SVM 和 FS_MLC 4 个模型进行最后分类结果的对比,并制作实验区土地利用分类图,如图 4 所示。对比图 4a、4b,经过特征选择后的随机森林算法依然能够较清晰地区分农耕区内各地物类型,相对于采用高维特征变量的分类方法,它能更高效快速地提取农耕区土地利用类型信息;对比图 4b、4c、4d,3 种分类方法中,在利用支持向量机和最大似然分类法进行信息提取时,错分现象明显,如图 4c 所示,多数库塘被错分为湖泊,图 4d 中,耕地被错分为水渠现象严重。综上所述,特征降维后,采用随机森林算法能够保证分类效果

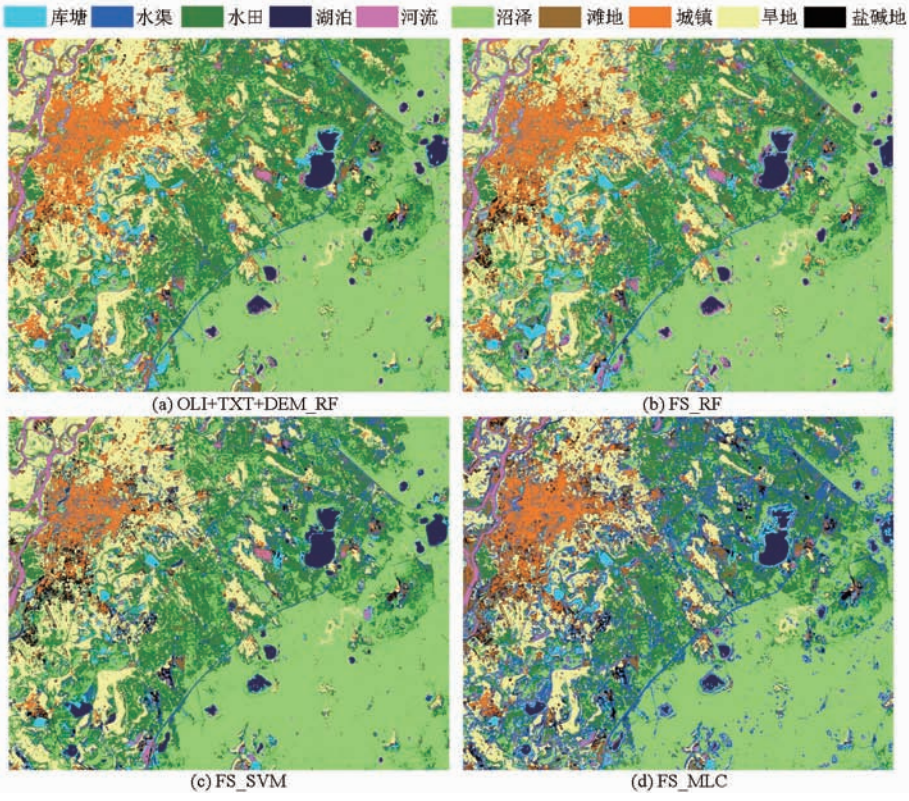

图 4 分类结果对比

Fig. 4 Comparison of different classification results

的同时,减少运作时间,真正实现了信息的高效提取。

5 结束语

基于随机森林算法,综合多季节、多时相光谱信息、纹理信息和地形信息,实现了对农业耕种区土地利用类型的划分。随机森林算法较传统分类方法可以有效利用样本间的差异,选择最优变量进行分类,其并行处理大量高维数据的能力显著提高了农耕区土地类型的识别效率。随机森林特征变量重要性可以作为特征选择的依据,有效降低数据维度,降维后的综合信息与随机森林算法的结合,真正实现了在分类效率、精度与适用性之间的良好平衡,为今后大面积提取农业耕种区土地利用信息提供了可靠方法和有利依据。

参 考 文 献

1 林楠,姜琦刚,杨佳佳,等. 基于资源一号 02C 高分辨率数据的农业区土地利用分类[J]. 农业机械学报,2015,46(1):278 – 284.
Lin Nan, Jiang Qigang, Yang Jiajia, et al. Classifications of agricultural land use based on high-spatial ZY1 – 02C remote sensing images[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015,46(1):278 – 284. (in Chinese)

2 李建刚. 不同土地分类标准协调研究[D]. 北京:中国地质大学,2012.
Li Jian'gang. Coordination of research in different of land use classification criteria[D]. Beijing: China University of Geosciences, 2012. (in Chinese)

3 魏继伟. 基于遥感图像的土地利用分类研究[D]. 长春:东北师范大学,2012.
Wei Jiwei. The study of land use classification based on remote sensing image[D]. Changchun: Northeast Normal University, 2012. (in Chinese)

4 孙丹峰,杨冀红,刘顺喜. 高分辨率遥感卫星影像在土地利用分类及其变化监测的应用研究[J]. 农业工程学报,2002,18(2):160 – 164.
Sun Danfeng, Yang Jihong, Liu Shunxi. Application of high-spatial IKONS remote sensing images in land use classification and change monitoring[J]. Transactions of the CSAE, 2002,18(2):160 – 164. (in Chinese)

5 买凯乐,张文辉. 黄土丘陵沟壑区遥感影像信息面向对象分类方法提取[J]. 农业机械学报,2011,42(4):153 – 158.
Mai Kaile, Zhang Wenhui. Object-oriented classification approach for remote sensing imagery information extraction in loess hilly-gully region[J]. Transactions of the Chinese Society for Agricultural Machinery,2011, 42(4):153 – 158. (in Chinese)

6 赵建华. 基于 SOM 神经网络的半监督分类算法[J]. 西华大学学报:自然科学版,2015,34(1):36 – 51.
Zhao Jianhua. Semi-supervised classification algorithm based on SOM neural network[J]. Journal of Xihua University: Natural

Science,2015,34(1):36 - 51. (in Chinese)

7 孟治国. BP 神经网络在土地利用分类中的应用分析[D]. 长春:吉林大学,2006.
Meng Zhiguo. The analysis of the application of the BP neural network in the land use classification[D]. Changchun: Jilin University, 2006. (in Chinese)

8 赵春晖,乔蕾. 基于改进的最小二乘支持向量机的高光谱遥感图像分类[J]. 应用科技,2008,35(1):44 - 52.
Zhao Chunhui, Qiao Lei. Classification of hyperspectral remote sensing image using improved LS - SVM[J]. Applied Science and Technology, 2008,35(1):44 - 52. (in Chinese)

9 傅文杰,洪金益,林明森. 基于光谱相似尺度的支持向量机遥感土地利用分类[J]. 遥感技术与应用,2006,21(1):25 - 30.
Fu Wenjie, Hong Jinyi, Lin Mingsen. A method of land use classification from remote sensing image based on support vector machines and spectral similarity scale[J]. Remote Sensing Technology and Application,2006,21(1):25 - 30. (in Chinese)

10 Jennifer M C, Joseph F K, Alisa L G. Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in northern Minnesota[J]. Remote Sensing,2013,5(7):3212 - 3238.

11 Rodriguez-Galiano V F, Abarca-Hernandez F, Ghimire B,et al. Incorporating spatial variability measures in land-cover classification using random forest[J]. Procedia Environment Sciences,2011(3):44 - 49.

12 Rodriguez-Galiano V F, Chica-Olmo M, Abarca-Hernandez F, et al. Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture[J]. Remote Sensing of Environment,2012,121:93 - 107.

13 Mellor A, Haywood A, Stone C, et al. The performance of random forests in an operational setting for large area sclerophyll forest classification[J]. Remote Sensing,2013,5(6):2838 - 2856.

14 Hayes M M, Miller S N, Murphy M A. High-resolution landcover classification using random forest[J]. Remote Sensing Letters, 2014,5(2):112 - 121.

15 王栋,岳彩荣,田传召,等. 基于随机森林的大姚县 TM 遥感影像分类研究[J]. 林业调查规划,2014,39(2):1 - 5.
Wang Dong, Yue Cairong, Tian Chuanzhao, et al. Classification of TM remote sensing image based on random forests of Dayao county[J]. Forest Inventory and Planning,2014,39(2):1 - 5. (in Chinese)

16 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报,2013,50(4):1190 - 1197.
Li Xinhai. Using random forest for classification and regression[J]. Chinese Journal of Applied Entomology ,2013,50(4):1190 - 1197. (in Chinese)

17 刘毅,杜培军,郑辉,等. 基于随机森林的国产小卫星遥感影像分类研究[J]. 测绘科学,2012,37(4):194 - 196.
Liu Yi, Du Peijun, Zheng Hui, et al. Classification of China small satellite remote sensing image based on random forest[J]. Science of Surveying and Mapping,2012,37(4):194 - 196. (in Chinese)

18 雷震. 随机森林及其在遥感影像处理中应用研究[D]. 上海:上海交通大学,2012.
Lei Zhen. Random forest and its application in remote sensing[D]. Shanghai: Shanghai Jiao Tong University,2012. (in Chinese)

19 黄衍,查伟雄. 随机森林与支持向量机分类性能比较[J]. 软件,2012,33(6):107 - 110.
Huang Yan, Zha Weixiong. Comparison on classification performance between random forests and support vector machine[J]. Software,2012,33(6):107 - 110. (in Chinese)

20 徐涵秋. 利用改进的归一化差异水体指数(MNDWI)提取水体信息的研究[J]. 遥感学报,2005,9(5):589 - 595.
Xu Hanqiu. A study on information extraction of water body with the modified normalized difference water index(MNDWI)[J]. Journal of Remote Sensing,2005,9(5):589 - 595. (in Chinese)

21 王书志,张建华,冯全. 基于纹理和颜色特征的甜瓜缺陷识别[J]. 农业机械学报,2011,42(3):175 - 179.
Wang Shuzhi, Zhang Jianhua, Feng Quan. Defect detection of muskmelon based on texture features and color features[J]. Transactions of the Chinese Society for Agricultural Machinery,2011,42(3):175 - 179. (in Chinese)

22 Breiman L. Random forest[J]. Machine Learning,2001,45(1):5 - 32.

23 Zhu Zhe, Curtis E W, John R, et al. Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat and SAR data[J]. Remote Sensing of Environment,2012,117:72 - 82.

24 Beijma S V, Comber A, Lamb A. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data[J]. Remote Sensing of Environment,2014,149:118 - 129.