CLASSIFICATION OF LAND TYPES IN MINERAL AREAS BASED ON CART

Wenbo Wu, Yuping Chen, Jiaojiao Meng, Tingjun Kang

School of Geomatics, Liaoning Technical University, Fuxin, Liaoning, China, chneyuping317@163.com

KEY WORDS: Remote sensing images, Knowledge, Knowledge based classification method, Classification and Regression Tree, Land types of the mineral areas

ABSTRACT:
The accurate classification of land types in mineral areas is very important to develop the mineral resources and monitor the environment of mineral area. Based on the spectral characters and the spatial knowledge of the ground objects in a certain mineral area in Shenyang which served as a training area. The spectral characters, texture characters, digital elevation model and slope are selected and extracted. The defined training sample areas are picked up by stratified random sampling techniques based on geographical coordinates. Firstly, using Classification and Regression Tree (CART) to discovery classification rules which integrates spectral, textural and the spatial distribution characters from these samples. Then, the interpretation was performed by a judgement based on these rules. Finally, the traditional supervised--- Maximum Likelihood Classifier (MLC) was performed to check the classification accuracies. The results have suggested that the accuracy of classification based on the CART is higher and it can obtain a lot of reasonable rules most quickly and effectively. From the highly accurate classification results with multitemporal Landsat TM images we can detect the change of water bodies, vegetation, road, mining area, residential area and so on. So, it can provide useful data for the development and protection of the mineral resources.

1. INTRODUCTION

The excessive development of mineral resources is having initiated the grave resource environment problem. The precise classification of mining land is essential to develop the mineral resources and monitor the environment of mineral area. At present, a lot of scholars inside and outside have conducted the research regarding this , there are two methods they always used: One is to make use of aerial photograph or high resolution satellite image to establish interpretation markt by lettering, logical analyzing and field inspecting on the same scale topographic map, then extract the types of terrain feature by visual interpretation. The classification accuracy of this method is high, but it wastes time and energy [1]. The other is to make use of the image features after transforming to build classification template for each surface feature, and then use the maximum likelihood classification to classify. The main alternation methods include Principal Component Anlysis(PCA), Tasseled Cap, Ratio Method and so on[2]. The classification accuracy of this method is not high , and it isn't applicable for distributed and scattered mining ares.

With commercial high resolution satellite appearing, we will also be confronted with especially austere challenge by handling a great deal of remote sensing data while enjoying the advance of science and technology[3]. How to extract what we need information and discover implicit knowledge from a mass of data is the key question of remote sensing image interpretation at present. Such, data mining and knowledge discovery technology which emerged in1980s have being introduced to the image processing of remote sensing, and the decision tree as a type of data mining technology was led into this field just 10 years ago.The decision tree method has many merits,such as building-up quickly, high accuracy, can creating understandable trules, the little amounts of calculation. It's introduction can make full use of remote sensing data ,so it has an advantage in information extraction.The principle of decision tree is not to try to use one kind of algorithm or a decision rule to classify many category at a time, but is to carry out the most effective classification specifically for different aggregation choosing the different standard or method, so it can oversimplify with complicated problems and resovle completely[4].

There are many ways to structure decision tree. The comparatively mature algorithms include ID3 , C4.5 , C5.0 series brought forward by Quinlan, Classification And Regression Tree (CART) , SLIQ and CHAID brought forward by Breiman and Friedman.The various algorithms have respective advantage and deficiency, and ID3, C4.5, C5.0 series are much uesed in remote sensing field, but these algorithms all adopt pruning-before, it needs to adjust the parameter carrying out trial again and again,so this research adopted CART pruning-after.

2. IMAGE PREPROCESSING

In the study,we chose TM image gained in 11th, August, 2001 of this area as data source.Chose the 1:50000 scale topographical map of the area as the reference coordinate, and then carried out geometric precise correction. The amount of residual deviation was less than 5m, and the error didn't exceed one pixel, this result satisfied the study. After the correction, the size of picture element was 30m, and then selected a trial area whose size was 256×256 pixel from it. Finally, digitized the contour lines on the 1:50，000 scale topographical map, the contour inteval was 10m, then made use of these contour lines to creat DEM of the trial area., and then created slope view and aspect view based on this DEM. The overlay of TM's false color composite imagery (RGB543) and DEM is as following figure 1.

Figure 1. The overlay of TM's false color composite imagery (RGB543) and DEM

3. THE PRINCIPLE OF CART

In fact, CART is one kind of data surveying and forecasting algorithm (Breiman L, 1984), it not only can deal with the highly tilt and many states numerical value,but also can deal with homothetic attribute data in order or out of order. The CART algorithm adopts the technology of the dimidiate recurrent division, it always divides the current sample into two son-samples, this makes each non-leaf nodes that has two branches.The benefit of this algorithm is that it can take a portion as the training data, and the other one as the checkout data.It leads into an "adjustable mistake rate " in process, it means that all leaf nodes of one branche joined a punishment factor.If that branch is still able to keep low mistake rate, then keep it,otherwise, give it up.The ultimate analysis result is an optimum binary decision tree which takes complicated degree and mistake rate into account, all approachs that equinoctial points define are corresponding to a most conditional class. So,the decision tree that CART creates is a concise binary decision tree.

The particular description of CART is as following:

/* T represents the current sample collection, $T\_attributelist$ represents the current candidated attribute collection */

Function cartformtree (T)

{

establishe root node N ;

assign classes for N ;

If T all belong to the same class OR only on sample left in T

Then return N as leaf node and assign a class for it ;

For attribute in each $T\_attributelist$

carry out a division for that attribute, calculate $Gini$ index of that division ;

the testing attribute of N equals to the attribute which has minimal $Gini$ index among $T\_attributelist$;

divide T into two son-collections T_1, T_2 :

transfer cartformtree (T_1)

transfer cartformtree (T_2)

}

CART has the following merits: limpid structure, easily understand,simply realize,quick speed, high accuracy; Can deal with a large amount of data and the nonlinearity relation.The data put in can be continuation variable also can be a discrete value; Contains the default and error of a data; Can give out the significance of the testing variable [5 , 6]. In the process of CART decision tree growth, it adopts $GiniIndex$ which is always used in economics field to be the criterion testing variable and segmentation rule.The mathematics definition of $GiniIndex$ is as following:

$$GiniIndex = 1 - \sum_{j}^{J} p^2(j \mid h)$$

$$p(j \mid h) = \frac{n_j(h)}{n(h)}, \sum_{j=1}^{J} p(j \mid h) = 1 \qquad (1)$$

Where, $p(j \mid h)$ is the probability when some one testing variable h belongs to the j class, $n_j(h)$ is the sample number when some one testing variable h belongs to the j class, $n(h)$ is the sample number when some one testing variable belongs to j , J is the number of class in training sample collection. [7]

4. CLASSIFYING BASED ON CART

4.1 The selection of training sample

The selection of training sample was the essential step in the study, which directly related the rule quality gained. There still weren't standard classification system for mining areas based on the remote sensing image, this article referred to the land use and land cover classification system.

In order to study the overall situation of the mining land resource, considering the actual situation, the trial area land types was classified into seven classes according to the image interpretation ability.The land types included Water body, Paddy field, Arid land, Building area, Road,Vegetation and Subsidence land.

In order to enable the training sample to reflect each kind of land type in the spatial distributed characteristic, this article used random delamination sampling method according to the space coordinates.Carried the sampling on the trial area referring to TM and 1:50,000 scale topographic map.

4.2 Determining the testing variable

The spectral response characteristic most direct affects the ground fetures identification ability of multispectral remote sensing image, and it is also the most important interpretation element.Each ground feature has the unique spectrum reflection and the radiation characteristic as a result of the different material composition and the structure, this reflects on the image is that each ground feature in various wave bands has different grey level. But because of the complexity of the ingredient and structure of ground features, as well as the influence of the remote sensing sensor and the atmospheric environment, the optical spectrum feature of the ground features present the multiple complex changes.Therefore, in order to make full use of the TM data to carry on the information

extraction and the classification, first must analyze the information characteristic earnestly.The grey level curve of different ground features on different wave bands is as following figure 2.

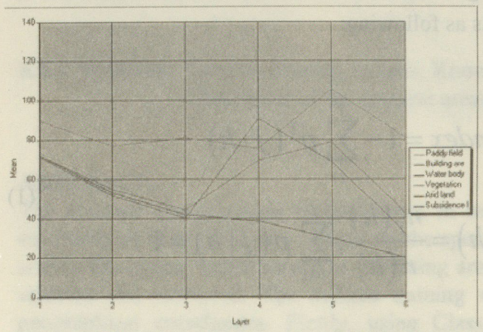

Figure 2. The grey level curve of different ground features on different wave bands

The texture also is the remote sensing image important information, which not only has reflected the image grey level statistics information, but also has reflected the ground feature itself structure characteristic and the spatial arrangement relations, so it also is one of the visual interpretation important symbols[8]. Many research indicated that, it can enhance the interpretation precision to overlay the texture information on the primary image spectrum information. In the remote sensing image, the texture mostly is random, which obediences statistical distribution, it often described by Grey level Co-occurrent Matrix. Grey level Co-occurrent Matrix is a matrix which is constituted by level-two union conditional probability density between the image grey levels, it has reflected the correlation between two random points grey level in the image.

In this foundation, we had defined eight texture characteristic statistical value: Mean, Variance, Homogeneity, Contrast, Dissimilarity, Entropy, Second Moment,all the value described the image texture characteristic from many aspects.[9] Using Grey level Co-occurrent Matrix to carry on the texture analysis involves three Important parameters: motion window size, motion step and motion direction. The appropriate window size is especially important to the texture analysis.In general, the window size is decided by the primary image texture structure, the small window denotes the slight texture characteristic while the big denotes the rough. The choice of the motion step is also decided by the image texture granularity, the short step is suitable for the slight texture while the long is suitable for the rough.Many researches pointed out taking the step as one is quite effective to all the different texture. In studies,we usually take four main direction values of 0° , 45° , 90° , 135° direction or the mean of these four direction values.[10] However, Franklin and the Pebble pointed out the texture characteristic of single direction was better than the mean of these four directions. [11] Through analyzing and comparing different window sizes and steps, the study selected the window whose size was 5×5 and the step was one.

In addition, because of the region natural condition control and the human factor intervention, the spatial distribution of the ground features often have some region differentiation rules.Therefore, this study also took the geographic coordinates and the terrain factor as the forecast variables.

In summary, the study used 18 following testing variables: two geographic coordinates (X, Y), six wave band grey levles of TM image(1,2,3,4,5,7 bands)(the 6th wave band exception), seven texture characteristics (Mean, Variance, Homogeneity, Contrast, Dissimilarity, Entropy, Second Moment (SM)and three terrain factors (DEM, Slope, Aspect) . Among these, the slope was separate variable whose value is 0,1,…,7, and the other were continual variables.The goal variablesof this study was: Water body, Paddy field, Arid land, Building area, Road,Vegetation and Subsidenceland.

4.3 Classifying based on CART

In the study ,we first chose 3,200 sample collections for these above-mentioned testing variables and target variables, then make use of CART to analyze and study the samples,finally,we structured a decision tree who had 38 leave nodes. Its studying accuracy is 92.8% and verification accuracy is 90.3%. The structure of the decision tree can be expressed into the *If-Then[CF]* form very conveniently. The testing route from the tree's father node to every leaf node corresponded to a rule , so there were 38 rules in all. For examples:

If (19.688< Mean<=19.958 && 10.000<TM6<=45.000 && 13<TM5<=71.000) Then (class = 1) CF=0.987

If (42.333<Mean<=44.500 && 40.000<TM6<=44.00 && 42.000<TM3<=47.000 && 4578510<X<4578522) Then (class = 7) CF=0.854

In the *If-Then[CF]* form, *CF* indicates the confidence measure of the rule,and its value region of *CF* is [0 , 1]. If the value is 0, then the possibility of the present pixel belongs to the given class out is also zero; while the value is 1, then the confidence value is invariable. Making use of the above-mentioned rules and the following simple matching strategies, we gained the classified image such as figure 3(a).

1. When only satisfied some one rule, then took the output category of the rule as the classified category.
2. When simultaneously satisfied multi-rules, then took the output category of the rule whose confidence value was more as the classified category.
3. When didn't satisfied all the rules, then it was not classified.

In order to compare with other classification methods, the study also had used supervised classification to classify and test, the classified image such as figure 3(b).

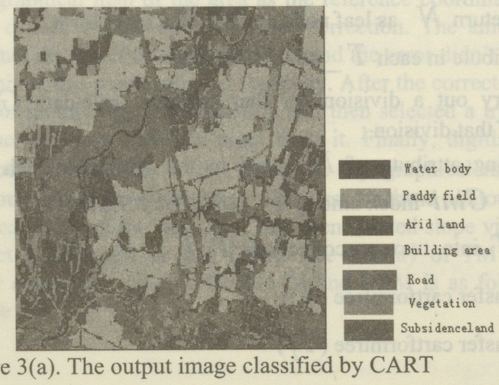

Figure 3(a). The output image classified by CART

Figure 3(b). The output image classified by Supervised Classification

4.4 Accuracy Assessment

Accuracy assessment is one of the indispensable jobs in the process of the remote sensing data classification.Through accuracy assessment, the classifying person can ascertain the validity of the classification signature, improve the classification signature, enhance classification accuracy; the user can gain information in the classified image correctly and effectively according to the classification accuracy. [12] The method based on the confusion matrix is universally recommended classification accuracy assessment method. In this study, we chose 1,024 detecting samples randomly referring to TM image and 1:50000 topographic maps, then through visual interpretation to structure the confusion matrix, carried out classification accuracy assessment based on the correlation index calculated.

5. CONCLUSIONS

This study used the thought of CART analysis whose tree-shaped was simple,clear and intuitionistic. It caused the multi-characteristic and multi-model land types of the trial area to be clearer, so it easily realized the automatic recognition in the compute.

This study used CART analysis to classify and extract the mining area land resources, had yielded some result, specially had extracted the subsidence lands.But,because of the time and insufficiently experienced myself, the CART decision tree was not too perfect. After ground truth investigation, we found one subsidence land was omited and two arid lands had classified into subsidence land by mistake, but this thought was feasible, in later work can improve the CART decision tree shape, cause it to be more perfect.

REFERENCES :

[1] Dazhi Guo, Yehua Sheng, Mingxing Hu etc., 1998. *Mining Area Environment Disaster Dynamic Monitor and Analysis Assessment*. The press of Chinese Mining University, Xuzhou.

[2] Yinhui Zhang, Gengxing Zhao, 2002a. The summary of remote sensing data classification methods about land use/ land cover. *Chinese agricultural resources and districts*, 23(3), pp. 21-25.

[3] Kaichang Di, 2001. *Spatial Data Mining and Knowledge Discovery*. The press of Wuhan University, Wuhan.

[4] Yurong Gao, 2006b. Study on land use information extraction based on decision tree method. *Dissertation Submitted to Zhejiang University For Degree of Master*, Zhejiang.

[5] Breiman L, Friedman J H, Olshen R A, etc., 1984. *Classification and Regression Tees*. Monterey, Califomia, U.S.A.: Wadsworth Intemational Group, pp. 1-358.

[6] Yohannes Y, Hoddinott J, 1999. *Classification and Regression Tree: An Introduction*. Washington, D.C., U.S.A.: International Food Policy Research Institute.

[7] Ping Zhao, 2003b. Knowledge-based Landuse/cover Classification in the Typical Testareas of the Lower Reaches of Yangtze River. *Dissertation Submitted to Nanjing University For Degree of Doctor*, Nanjing.

[8] Haralic R M, Shanmugam K, 1973a. Dinstein I Texture Features for Image Classification. *IEEE Tmnsactions on Systems, Man and Cybemetics*, (6), pp. 610-621.

[9] Deshen Xia, Desheng Fu, 1997. *The Modern Image ProcessingTtechnology and Application*. The press of Southeast University, Nanjing.

[10] Treitz P, Howarth P, 2000a. Integrating Spectral Spatial and Terrain Variables for Forest Ecosystem Classification. *Photogrammetric Engineering & Remote Sensing*, 66(3), pp. 305- 317.

[11] Franklin S E, Pebble D R, 1989a. Spectral Texture for Improved Class Discrimination in Complex Terrain. *Intemational Joumal of Remote Sensing*, 54, pp. 1727- 1734.

[12] Jianping Wu, Xingwei Yang, 1995a. Accuracy analysis of classification of remote sensing data. *Remote Sensing Technology and Application*, 10(1), pp. 17-24.