GENERATION OF KNOWLEDGE RULES ON THE EXTRACTION OF COASTAL WETLAND

ZHANG Yue [a,b], RUAN Renzong [a], DING Xianrong [a], WANG Weiping [a], XU Hui [a], GE Xiaoping [a], ZHANG Xiaoxiang [a], XU Jun [a]

[a] State Key Laboratory of Hydrology, Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098
[b] College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China
- lisa_ling7892002@hhu.edu.cn

KEY WORDS: Land Use, Data Mining, Multispectral, Knowledge Base, Mapping, Landsat

ABSTRACT:

The wetland is one of the most important ecological systems. The present condition and its change trend of wetland are very important to the making of policy upon the reclamation, exploitation, management and protecti on of wetland. In this paper, the coastal wetlands in the northern part of Jiangsu Province are taken as study object and the technology about the extraction of wetland is explored by using multi-features and knowledge rules and multi-spectral Landsat 7 ETM+ acquired on May 26, 2005, in combination with the analysis upon the characteristics of wetlands and its presentation in remotely sensed imagery and data of field investigation of the same period. Based on the analysis of the characteristics of spectrum about the wetland, in this paper, at first, unsupervised classification on the image of study area was conducted. And then, by using the spectral feature of wetlands, texture, principal component analysis, NDWI and relative knowledge rules, the results of unsupervised classification was improved. The accuracy of extraction of wetlands has been greatly improved, from 71.09% to 87.16% and KAPPA coefficient from 0.6546 to 0.8438. The results showed that the classification accuracy of the extraction of wetland using knowledge rules has a very great improvement. By the support of remote sensing software such as ERDAS, ArcGIS and S-PLUS, we have improved classification accuracy of wetland at 72.76%. By comparison, the classification accuracy of extraction by using knowledge rules on the results of unsupervised classification is very high.

1. INTRODUCTION

Wetlands are valuable ecosystems that play important roles in our environment. It is human settlement and natural resources, and also one of the ecological landscapes full of biodiversity in the nature.

Wetlands are always the research focus and important field of the researchers in geography, biology and other subjects. Generally speaking, wetlands are lands where saturation with water is the dominant factor determining the nature of soil development and the types of plant and animal communities living in the soil and on its surface (Cowardin et al., 1979; Lyon, 1993). Wetlands are valued for their ability to store floodwaters, protect shorelines, improve water quality, and recharge groundwater aquifers (Daily 1997). In many countries, local economies depend on wetlands for fisheries, reed harvesting, grazing, and recreation. Wetlands are related to the humankind closely.

The classification of remotely sensed image is the one of the most important links in the research of land use change and also the basis of the extraction of the land use change information (Zhang X., 1997; Chen B., 2002; Li S., 2002). How to improve the classification accuracy of remotely sensed image has been and now still is one of hotspots on remotely sensed researches. A lot of scholars home and abroad have carried out much positive and useful exploration (Li S., Yun C., 1999; Li X., 1995; Gan F. et al., 1999). At present, various methods have been developed and practiced for land-use/land-cover. There are several major approaches to these methods: multi-source, multi-date remotely sensed data merge classification method; remotely sensed image classification supported by GIS method; non-parametric remotely sensed image classification method. Recently, The non-parametric classification methods different from conventional models such as Artificial Neural Network (ANN), intellectual technology, vague mathematics, expert system and so on have matured step by step, available for application continually. Some achievements have been got on the aspect of wetland research, such as Carpenter use the remotely sensed imagery of Landsat for modeling to monitor inland water quality of the wetland (Carpenter D. J., Carpenter S. M., 1983). Wani execute remotely sensed quantitative study of suspended sediment concentration for India Dal Lake by the use of IRS LISS-II remotely sensed data (Mcfeeters S. K., 1986).

Based on wetland features and attributes of remotely sensed image, by the analysis of wetland multi-spectral remotely sensed image features, the paper explores the regulations and methods of the extraction of wetland so that we can provide foundations for the protection and development of wetland.

2. STUDY AREA AND DATA SOURCES

Jiangsu Province has the shoreline with the length of 954 km and the sea area encompasses 3.75 km^2. Jiangsu Province is one of the provinces with rich wetland resources (Wang J., 1999). Its total area is of 4365000 hm^2 taking 42.5% of the total territory area of the province.

A combination of data from surveying and remote sensing data were used in this study. Data from Landsat-7 Enhanced Thematic Mapper Plus (ETM+) bands 1-5 and 7 were extracted from imagery acquired on May 26, 2002 (Figure 1). In addition, we also used the investigation data of coast from SEA 908.In Figure 1, the red lines stand for the coast investigation line of

SEA 908, and the red points stand for the sample points in the investigation.

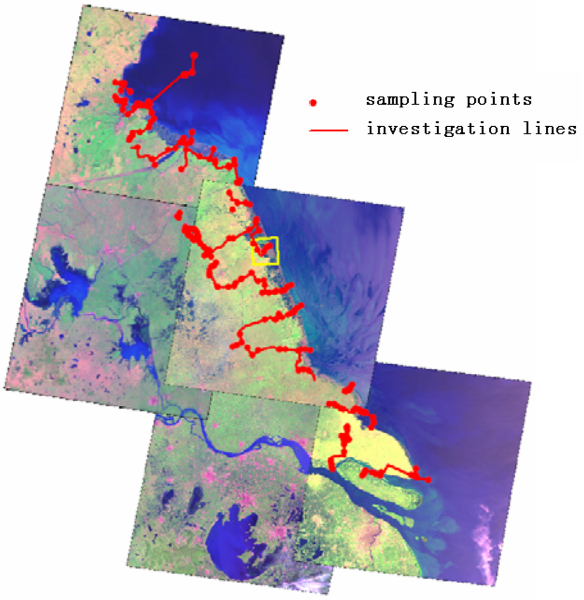

Figure 1. Location of Study area

Seven land cover types were recognized in the study area. The classification scheme of this paper is listed in Table 1:

ID	Classification Name	Description
1	spartina	
2	salsa	
3	reed	
4	paddy fields	
5	dry land	Using for cropping
6	aquaculture	Using for cultivation
7	water	Sea and rivers

Table 1. Classification system for the study area

3. STUDY METHODS

3.1 The analysis on spectral features of wetland

Having the same conditions such as structure features, illumination and so on, homogeneous objects usually have same or similar spectral features so that they show some intrinsic similarities, theoretically being assigned to the same spectral space region; different objects, having different spectral features, are assigned to different spectral space region. However, due to the variety of environment and seasonal effect, different objects having the same spectral features and the same objects having different spectral features extensively exist. Therefore it is difficult to classify objects based solely on spectral characteristics. In some bands, different objects are similar to each other sometimes, so we should explore and use other features to discriminate the objects.

As shown in Figure 2, the spectral curve of spartina and of salsa both have a valley in TM4 and a peak in TM5. Thus, we can use the ratio index of the two bands as the discriminative factor to separate them from other objects. However, the spectral mean of salsa is higher than that of spartina in TM3. The reason is because in most cases spartina grows in the water while salsa grows in a more dry condition, the spectral value of salsa is a bit higher than that of spartina, especially in TM3. Reed also belongs to vegetation, but its spectral mean curve has lots of differences from the former two. The curve is similar to the water, but it doesn't have obvious downtrend in TM4. Reed grows in the water, and the water level is high when the image is acquired, so the spectral feature is influenced by the water partially. At the same time, it is also influenced by chlorophyll, so it hasn't obvious uplift in infra-red, but is much higher than that of water in the band. Paddy fields have the high spectral value in TM4 and TM5.This is because paddy rice is denser than the former three kinds of vegetation, its reflectivity is greater than in infra-red band. The spectral curve of dry land is similar to that of the soil in general, but having a valley value in TM4, possibly owning to the higher content of water or humus matter in the soil. Generally, the spectral curve of water and aquaculture is similar very much. However due to the effect of much phycophyta, water has higher spectral response in TM1 through TM3 than aquaculture. The phycophyta could absorb partially visible light, so that the reflectivity of aquaculture is lower than water.

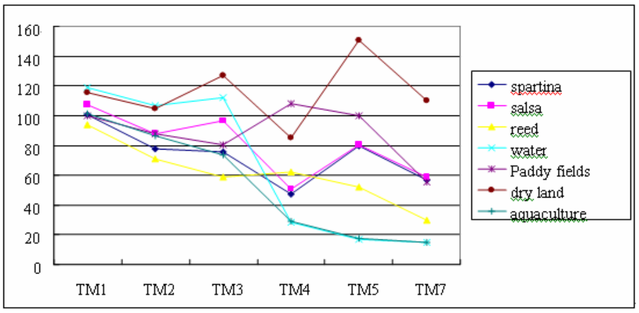

Figure 2. Statistical characteristics of spectral response in the study area

3.2 Unsupervised classification

The analysis technique of Iterative Self Organizing Data (ISODATA) of unsupervised classification is used in this paper to classify the image. Unsupervised classification techniques have been very popular for wetland classification since its suitability in dealing with objects with spectral variability and gradual transitions between vegetation types. In the paper, the image was classified into 70 clusters, which were then allocated among the 7 classes by using visual examination and ground truth data.

3.3 Logic identification based on knowledge rules

3.3.1 The feature variables used in the logic

The initial unsupervised classification results still have obvious classification error. This shows that the feature space of the six bands of TM image couldn't discriminate these objects well. To overcome these shortcomings, this paper exploits other methods such as the relation of feature variables. For example, we can do some transformation on the TM image, such as K-L transformation, tasseled-cap transformation, etc to further derive some other attribute variables of the objects, we can also find some index such as normalized difference vegetation index(NDVI),normalized difference water index(NDWI),etc, for further improving the classification accuracy by refining optimization of unsupervised classification results. This paper used the layered classification method to realize it.

3.3.2 The analysis of wetland spectral ratio: By the analysis of unsupervised classification result and the corresponding original image, we find out that although the confused objects have similarities of spectral response features in many bands, their change trends at some bands is discrepant, as the spectral curve rises and falls in different range. It provides a scientific base for the discrimination of objects using spectral ratio of two bands.

3.3.3 Texture analysis: Texture refers to the structure influenced by the regulated change of the targets inner color tone in remote sensing imagery; it reflects the degree of the image roughness. In the paper, when the spectral features are difficult to discriminate some classifications, the introduction of texture features will make it possible to discriminate them. The description methods of texture features include Gray Level Cooccurrence Matrix method, Autocorrelation Transformation method, statistical parameters method, digital transformation, etc (Wani, M. M., etc).

In the paper, the variance of 3×3 window size was used as texture metric, which described the variety of DN values between the pixel and around neighborhood pixels in each band. Under the same condition, the same kind of the object mostly have the similar texture features and the objects with larger difference usually belong to different kinds of objects. On this basis, generally speaking, if one region has the common variance characteristics, then they belong to the same kind of object to some degree.

3.3.4 PCA transformation: In the research, it was found that some street trees are confused with spartina. In order to explore other methods to discriminate them, Principal Component Analysis was used. A Principal Component Analysis was first conducted upon imagery, and then, to be convenient for analysis, we did linear extension on the transformed image. After linear extension, the range of the image pixel value was from 0 to 255.

3.3.5 Tasseled Cap transformation: In the above classification result, we noticed a high misclassification of reed objects to aquaculture. Under some conditions reed grow very sparse, this leading to the confusion with aquaculture in classification. However, because the reed belongs to vegetation while aquaculture is characteristic of water, certain differences exist between them in the aspects of brightness, greenness and moisture. Thus, in this paper tasseled cap transformation was used for the discrimination of them.

3.3.6 Normalized Difference Water Index: To exploit the differences in the reflectance patterns of water, water indexes are derived based on ratios or linear combinations of spectral responses in specific portions of the electromagnetic spectrum. In consideration of the obvious reflectivity difference of water in bands of green and infra-red, Mcfeeters (Zhao Y., 2003) proposed NDWI (Normalized Difference Water Index). NDWI is computed from near-infrared (NIR) and short-wave infrared (SWIR) as follows:

$$NDWI = (Green - NIR)/(Green + NIR) \qquad (1)$$

Where NDWI denotes Normalized Difference Water Index, Green represents reflectance in Green band, NIR reflectance in near infrared band. For the imagery of Landsat TM, band 2 is Green band and band 4 NIR band. Due to the reflectivity of water is gradually decrease from visible light band to middle infrared band, due to the strong absorbability in the range of near-infrared and middle infrared band, almost having no reflection, so that it is available to outstanding the water information of the image by the difference between visible light's spectral signatures and near-infrared's spectral signatures.

3.4 Classification based on machine learning

3.4.1 Theoretical basis: In the research, data mining was also used for the generation of knowledge rules for layered classification. The principles are as follows:

In a decision tree classifier, it is assumed that the feature variables and the target variable for a sample of pixels (training data), also known as labeled samples, are given. The class of the sample points has already acquired by the combination of visible interpretation and ground truth data. The sample points are divided into two parts, one for training and one for validation and pruning of the tree. The decision tree was then transformed into production rules for the later knowledge rule classification. The scheme is shown in Figure 3.

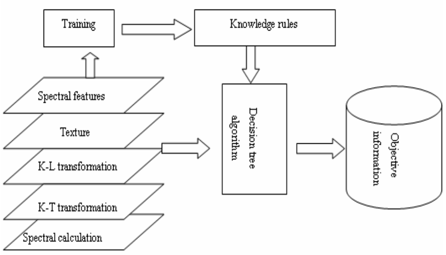

Figure 3. Flow chart of information of machine learning classification based on multi-feature set

3.4.2 Automatic generation of knowledge rules and classification: In this paper, we used classification and regression tree (CART) algorithm. The result of the CART analysis is a dichotomous decision or classification tree. Each path through the tree, defined by a series of dichotomous splits, specifies the conditions that lead to a most probable class. A part of tree is shown in Figure 4.

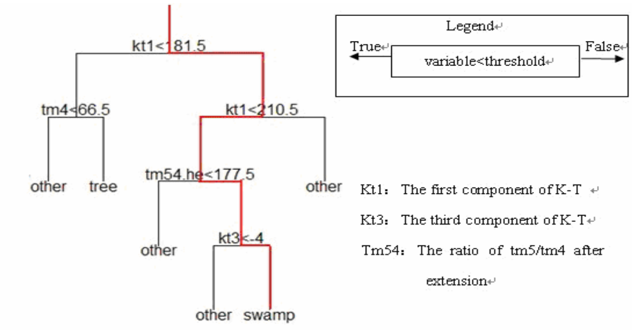

Figure 4. Knowledge Tree of S-PLUS

Then the decision tree was transformed into production rules. Taking the red branch of the tree in Fig 4 as an example, the production rule is as the following:

If（kt1>=182 && kt1<=210 && TM54.he>177 && kt3>=-3 && kt2>-91）Then（class = swamp2）

In this paper, the strategies for layered classification are as follows:

1. If only one rule will be met, the class value of the pixel will be the result of rule.
2. If several rules are met at the same time, the output class will be the result of the rule with the maximum confidence.
3. If no rule will be available, then the result will be “un-classified” class.

4. RESULTS AND ANALYSIS

4.1 Accuracy evaluation of unsupervised classification

The coastal wetland map was produced using an unsupervised classification procedure. And the overall accuracy was 71.09%, kappa coefficient was 0.6546.

The result shows that unsupervised classification method can identify and classify the objects of the study area to a certain degree of accuracy. However after classification some errors were easily identified where cover types were mapped out of their normal context. For example, the classification type presented high confusion between dry land and spartina, between reed and aquaculture, between water and aquaculture, and between reed and paddy fields. Thus we need to find out other methods for filtering out the misclassification wetland classes.

4.2 Inferring results of knowledge rules and analysis

After unsupervised classification, it is discovered that there are extensive confusion among spartina, salsa and reed. Due to the three kinds of objects all belong to vegetation and existing similarities of spectral reflectance, so we could not find out the difference among the three by the analysis of original TM image only. This paper collects samples on the three objects, obtaining the spectral mean values of the three at six bands of the TM image, plotting curve as Figure 3. We can see that between TM2 and TM3, it is discovered that the slope of spartina and reed curve are both less than 1, while the slope of salsa curve is beyond 1; similarly, between TM5 and TM6, the slope of spartina and salsa curve are both beyond 1, and the slope of reed curve is less than 1. According to the characteristics, the ratio indices TM2/TM3 and TM4/TM5 had been found useful for the discrimination of the three objects from each other.

In order to separate paddy from spartina and salsa, after getting the texture image, we collected samples on the three objects and plotted the curve with the sample means.The three kinds of objects are very different from each other, so that we select 10 as the threshold to separate paddy fields from spartina and salsa. This may indicate that in the case of lack of spectral differences between vegetation communities, in some periods of the growing season, textural information can be used for discrimination.

The features of the image after a tasseled cap transformation respectively concentrate on brightness, greenness and moisture. We can use it to discriminate the objects having difference on the above features. For example, in this paper, we used the method to separate the reed misclassified from aquaculture. After sample collection, obtaining the means of the samples, at last, we selected 115 in the sixth component as the threshold for discrimination between them.

Pre and post conduction of aquaculture exaction results are as follows:

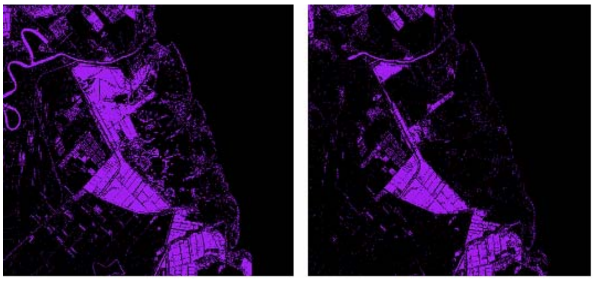

Figure 5. Extraction result of farm before and after processing

From Figure 5, it can be seen that the most part of reed are separated. Although a part of river is also separated from the class, it can be improved by post-classification processing.

Now we can build a decision tree using the knowledge rules produced by the above ratio index, texture analysis, K-L transformation, NDWI.

Then, we introduce the classification procedure using spartina as example. First, selecting 10 at the fourth band of texture image as the threshold, if the texture value at the fourth band is less than 10, it is the class contained spartina; if the value is beyond 10, we should exact spartina subsequently. And then, selecting 110 at the third component after K-L transformation, if the value at the third component is beyond 110, it belongs to spartina; if the value is less than 110, it belongs to street tree or other objects.

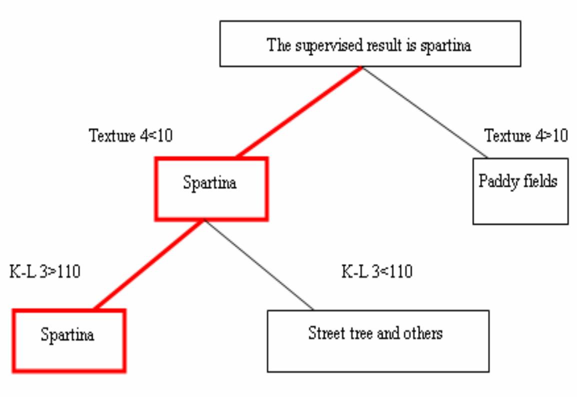

Figure 6. Knowledge rules on the extraction of spartina

The red branch in Figure 6 stands for the step-by-step extraction of spartina, separating other confused objects. According to the methods above, building decision trees for other classes respectively, we will get the optimum classification result. After refining process, the last classification results are as Figure 7:

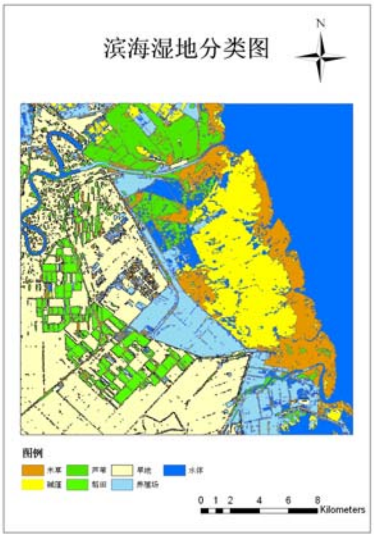

Figure 7. Extraction of wetland based on knowledge rule

From Figure 7 it can be seen that the thematic map becomes accurate than unsupervised classification result obviously and the classification result of each object conforms to the true situation grossly. After the knowledge rules based classification, the overall accuracy increased to 87.16% and Kappa coefficient to 0.8438.

Generally speaking, by applying knowledge rules, the classification accuracy was improved significantly, with the realization of effective separation of the objects. However, the classification yielded poor results for spartina. Since the plant was easily separated from other objects in spectral features and it was difficult to reach a good discrimination on other features, the accuracy of spartina was decreased. But the accuracy of the most objects was increased in some extent. Thus, the results indicate that the method of information extraction based on knowledge rules has higher practical value.

4.3 Classification based on machine learning and assessment

Decision trees on the extraction of wetlands was transformed into production rules and realized in Knowledge Engineer tool of ERDAS 9.0. After noise-removal processing, we evaluated the accuracy of classification result after noise-removal processing. The sample points used in accuracy assessment were still the ones used in the above two, and the evaluation results of classification accuracy are as follows:

Overall accuracy=72.76%, Kappa coefficient=0.6728.

The layered classification method based on automatic data mining can also reach relative high accuracy, with the prerequisite that we have accurate sample points. The sample number should reach a certain requirement and the sample points should distribute in each class uniformly. From the accuracy assessment result, it can be seen that the classification yielded good results for dry land, moderate results for water, and poor results for spartina and aquaculture. Since dry land have relative concentrated distribution and less confusion with other objects. But a part of pixels were also classified into the classes of reed and paddy fields, because there were footstalk distributed on the boundary of reed and paddy fields; the water in the study area is clear relatively and the content of suspended substance is lower. Among them, parts of pixels were classified as the dry land, with the reason that these are spatially close or intermixed with each other; relatively speaking, spartina and aquaculture were easily confused with other objects. A large part of pixels of spartina were classified into the classes of salsa, reed and paddy fields wrongly, for the reason that they all belong to vegetation, so there were great similarities in spectral features. A part of pixels of aquaculture were divided into water and dry land, with the reason that the spectrum and texture features of parts of aquaculture was similar to water, leading to difficult discrimination between them. Meanwhile, since satellite sensors such as Landsat TM has spatial resolution of 30 meter, often mixed pixels exist along the boundary of aquaculture and dry land, leading to a large confusion of the three; moreover, the accuracy of paddy fields was also not very high, mainly because its features was similar to salsa and reed and mixed pixels exist, causing difficulty in the generation of classification rules.

4.4 Comparison of classification accuracies of different methods

In the research, a same sample set was used to evaluate the accuracies of the three classification results, so that we can compare the qualities of different results objectively, as shown in Figure 8:

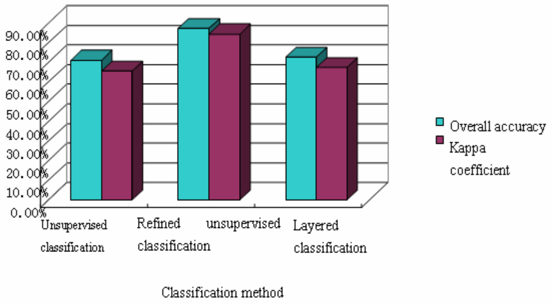

Figure 8. Comparison of accuracy of classification by the three methods

From Figure 8, the accuracies of the other two results were compared with that obtained from the results of spectral classification. The comparison showed that refined unsupervised classification produced the best result. The unsupervised classification itself already has quite a high accuracy because of multiple iterations. On this basis, the knowledge rules was generated for the purpose of further elimination of the errors of unsupervised classification effectively and improve the classification accuracy. On average the classification accuracy of the rule-based method was 87.16% compared to 76.19% for unsupervised classification. Towards the important classes, the classification accuracy also has great improvements: The class of aquaculture improved from 76.19% to 93.10%; the class of reed improved from 26.32% to 91.67%; the class of salsa improved from 80% to 89.29%. All the improvement is from the knowledge rules which can pointedly extract the objects. Moreover, having the sample points of high quality, the layered classification method based on data mining can reach a relative high accuracy identically.

5. DISCUSSION AND CONCLUSIONS

The study indicates that unsupervised classification can reach a certain accuracy requirements and acquire better classification result. However, if higher requirement is required, we should explore suitable knowledge rules to improve unsupervised classification result, making further optimization, making it close to the true condition of nature; to the classification based on knowledge rules, the most important is to obtain relative accurate rules. First, we should learn the relationships between potential indicators and object classes. We learn the growth rates of crop types, the optimal environmental conditions for crop growth in the various climatic zones, and the effects of environmental change on crop growth rate and quality. In addition, field survey data are used to learn how growth rates and crop quality are reflected in the available imagery data. And then, we can find out the feature vectors reflecting their difference mostly as the discrimination basis; moreover, we make full use of multi-source data. For example, we can use the topographic data in the study area, using GIS to produce DEM, and then, using slope, aspect from DEM and other variables as ancillary data to perfect the classification results. To the knowledge rules extraction based on data mining, we should pay attention to the selection of sample points. First, the samples number should reach a certain requirement and the number of sample for each kind should reach a certain standard. Second, we also should pay attention to the quality of the samples. That is, the sample points should be selected on the relative pure region to ensure the representative of the sample points. When the counts and quality are both reach the requirement, we can obtain relative high classification accuracy identically. An important advantage of the method based on data mining is that it can improve the accuracies of objects of interest accordingly by sacrificing the accuracies of other objects, getting the result of higher image classification accuracy.

For more effective use of satellite remote sensing, we should be aware of the limitations and advantages of satellite data and should choose data from their available wetland mapping options accordingly. Techniques for improving the classification of wetlands with satellite remote sensing data include the use of multi-temporal imagery and ancillary data. Multi-temporal imagery allows for the highest accuracies in wetland identification and discrimination from other land cover types. Layered or rule-based methods generally provide better results than conventional statistical classification methods, often because of their use of ancillary data. Moreover, radar is useful for studying wetlands because of its ability to distinguish between flooded and non-flooded areas, even in forests. In addition, radar data can be collected in almost all weather conditions, a characteristic that is especially important in areas with frequent cloud cover. In conclusion, satisfactory results will require specialized techniques.

ACKNOWLEDGEMENTS

This research was funded by a Natural Science Foundation 40771137; SEA 908 for providing coastal images of Jiangsu Province.

REFERENCES

Chen B.,2002. Multi Temporal Satellite Remote Sensing Data Application for Dynamic Inspection and Analysis of urban sprawl and environment change in Xiamen city. *Master degree thesis of Fuzhou University*, pp. 23-37.

Carpenter D. J., Carpenter S. M.,1983. Modeling inland water quality using Landsat data. *Remote Sensing of Environment*, 13, pp. 345-352.

Gan F., Wang R., Wang Y., etc,1999. A Study of Remote Sensing - based Classification for LUCC (Land Use/Cover Classification). *Remote Sensing for Land Resources*, 4, pp. 40-44.

Li S., Ding S., Xu S.,2002. Research in Method of Remote Sensing Image Classification. *Journal of Henan University*, 32(2), pp. 72-74.

Li S., Yun C.,1999. Study on Layered Methods and Its Application for Extracting Land Cover Thematic Information. *Remote Sensing Technology and Application*, 14(4), pp. 23-27.

Li X.,1995. A New Method to Improve Classification Accuracy with shape Information. *Journal of Remote Sensing*, 10(4), pp. 279-286.

Mcfeeters, S. K.,1996. The Use of Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features. *International Journal of Remote Sensing*, 17(7), pp. 1425-1432.

Wani, M. M., Choubey, V. K., and Joshi, H.,2000. Quanti cationof suspended solids in Dal Lake, Srinagar using Remote Sensing Technology. *Journal of Indian Society of Remote Sensing*, 24, pp. 25–32.

Zhang X., Huang Z., Zhao Y.,1997. *Remote Sensing Digital Image Analysis*, pp. 243-265.

Zhao Y.,2003. *Analysis Principle and Methodology of Remote Sensing Application*, pp. 414-424.