LAND USE CLASSIFICATION OF REMOTE SENSING IMAGE WITH GIS DATA BASED ON SPATIAL DATA MINING TECHNIQUES

Deren LI , Kaichang DI, Deyi LI*
(School of Information Engineering, Wuhan Technical University of Surveying and mapping,
No. 129 Luoyu Road, Wuhan, P. R. China, 430079)
(*Institute of China Electronic System Engineering, No.6, Wanshou Road, Beijing, P. R. China, 100036)
Email: dli@dns.wtusm.edu.cn kcdi@public3.bta.nat.cn

Key Words: Data Mining; Knowledge Discovery, Land Use Classification, Inductive Learning, Learning Granularity

ABSTRACT

Data mining techniques are studied to discover knowledge from GIS database and remote sensing image data in order to improve land use classification. Two learning granularities are proposed for inductive learning from spatial data, one is spatial object granularity, the other is pixel granularity. The characteristics and application scope of the two granularities are discussed. We also present an approach to combine inductive learning with conventional image classification methods, which selects class probability of Bayes classification as learning attributes. A land use classification experiment is performed in the Beijing area using SPOT multi-spectral image and GIS data. Rules about spatial distribution patterns and shape features are discovered by C5.0 inductive learning algorithm and then the image is reclassified by deductive reasoning. Comparing with the result produced only by Bayes classification, the overall accuracy increased 11 percent and the accuracy of some classes, such as garden and forest, increased about 30 percent. The results indicate that inductive learning can resolve the problem of spectral confusion to a great extent. Combining Bayes method with inductive learning not only improves classification accuracy greatly, but also extends the classification by subdivide some classes with the discovered knowledge.

1 INTRODUCTION

The integration of remote sensing and GIS is a topic of general interest in the field of photogrammetry, remote sensing and GIS. It is mainly contributes to two kinds of applications. One is GIS database updating by remote sensing images, the other is remote sensing analysis by the support of GIS data. These two aspects complement each other to make the GIS databases updated continually.

It has been long acknowledged that GIS data can be used as auxiliary information to improve remote sensing image classification. In previous studies, GIS data were often used in training area selection and post processing of classification result or acted as additional bands. Generally, it is accomplished in a statistical or interactive manner, so that it is difficult to use the auxiliary data automatically and intelligently. If the classifier does not request that the data have certain statistical characteristic, it is a simple and feasible way to use the auxiliary data as additional bands. But if the classifier requests certain statistical characteristics, the additional band method can not be used because most auxiliary data do not meet the requirements of statistical characteristics.

On the other hand, expert system techniques were incorporated in remote sensing image classification to make use of domain knowledge and logical reasoning. But building an expert system was very difficult because of the “knowledge acquisition bottleneck”. The traditional way of knowledge acquisition is that the knowledge engineer talks with the domain expert and then represents and inputs to computer in a formal format. This is usually a long and repeated process that can not avoid missing of information. Consequently, it is very difficult to put an expert system into practical use in remote sensing image classification.

In fact, large amounts of knowledge that can be used in image classification are hidden in GIS databases. Some knowledge is “shadow”, which can be extracted by GIS query. For example, “Is there any river in a area?”, “What is the maximum and minimum width of the roads?”, and so on. Some other knowledge is “deep”, such as spatial distribution rules, spatial association rules, shape discriminate rules, etc., that is not stored explicitly in the database but can be mined by computation and learning.

Spatial data mining and knowledge discovery (SDMKD), is the extraction of implicit, interesting spatial or non-spatial patterns and general characteristics. In [Li, et al., 1997], we proposed a theoretical and technical framework of spatial data mining and knowledge discovery. And spatial data mining is supposed to be used in two aspects, one is intelligent

This research was supported by Ph.D. program foundation from Ministry of Education of China and research grant (WKL(97)0302) form National Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing

analysis of GIS data, the other is to support knowledge driven interpretation and analysis of remote sensing images. SDMKD provides a new way of knowledge acquisition for remote sensing image classification. Some researchers have done valuable work in this field. Eklund et al. extracted knowledge from TM images and geographic data in soil salinity analysis using inductive learning algorithm C4.5 [Eklund et al., 1998], Huang et al. extracted knowledge from GIS data and SPOT multispectral image in wetland classification using C4.5 too [Huang et al., 1997]. In these two studies, geographic data were converted from vector to raster format in which the sampling size is equal to image pixel size. The implementation of data mining techniques in spatial database, especially inductive learning method, and the combination or integration of inductive learning with traditional image classification methods, are still need to be further studied.

In this paper, data mining techniques are studied to discover knowledge from GIS database and remote sensing data in order to improve land use classification of images. The paper is organized as follows. Section 2 describes the implement of inductive learning in spatial databases. Section 3 presents the methods of inductive learning in remote sensing image classification. Section 4 describes an experiment of land use classification of SPOT multispectral image. Finally we come to a conclusion.

2 INDUCTIVE LEARNING AND ITS INPLEMENT IN SPATIAL DATABASE

There are a lot of methods can be used in spatial data mining [Li, et al., 1997], among them inductive learning is a most import one. And there are many inductive learning algorithms which mainly come from the field of machine learning, for example, AQ11 and AQ15 by Michalski, AE1 and AE9 by Jiarong Hong, CLS by Hunt, ID3, C4.5 and C5.0 by Quinlan, CN2 by Clark, etc [Hong, 1997]. ID3 series, including ID3, C4.5 and C5.0, are most famous and influential.

ID3, which is a kind of decision tree algorithm, adopts a strategy of “divide and conquer”. It selects classification attributes recursively based on information entropy [Quinlan, 1993]. ID3 runs fast in learning and classification, this makes it effective for large database. The shortcoming of ID3 is that the decision tree is not clear as production rules, especially when a decision tree is large, it is very difficult to understand what does the tree mean. The other shortcoming is that ID3 can only deal with discrete attributes and it is restricted to two-class problems. C4.5, which is a extension of ID3, can covert a decision tree to equivalent production rules and can deal with multi-class problem with continuous attributes. These new features make C4.5 practical and most popular in the field of artificial intelligence and machine learning. C5.0 is a further improved version of C4.5, which runs much faster in very large databases. Therefore, we study the implementation of inductive learning in spatial database using C5.0 algorithm.

C5.0, as many other inductive learning algorithms, require that the training data are composed of several tuples and each tuple has several attributes one of which is class label. If we treat records as tuples and fields as attributes, these algorithms are very suitable for learning in relational database. Spatial data structure is more complex than the tables in ordinary relational database. Besides tabular data, there are vector and raster graphic data in spatial database. And generally, the features of graphic data are not explicitly stored in the database. Therefore, learning in spatial database is more difficult than learning in ordinary relational database in selecting the tuple and attributes of training data.

We regard learning tuple selection as a problem of determining learning granularity. Two learning granularities are proposed for inductive learning from spatial data, one is spatial object granularity, the other is pixel granularity. Spatial object represents area, line and point objects in graphical database or area and linear features extracted from remote sensing images. Pixel simply means the pixels of remote sensing images or cells of raster graphic data. Learning in spatial object granularity can discover knowledge concerning location, shape, spatial relation, etc. The discovered knowledge is generalized and can be used in intelligent spatial data analysis and also in remote sensing image classification. When the discovered rules are pplied to image classification, the image must be clustered or pre-classified to area or linear features before the rules are used. Learning in pixel granularity, on the other hand, can discover knowledge about spectral, location, elevation, etc. The discovered rules are more specialized and suitable for image classification, but not suitable for spatial data analysis and decision support. The two kinds of granularities have their own shortcomings as well. Learning in pixel granularity can not utilize shape information and it is difficult to utilize spatial association information. Learning in spatial object granularity can not utilize the detail information within the object, for example learning in polygon granularity can not utilize the accurate elevation and slope value within a polygon, and can only use a average or sample value. These two kinds of granularities should be selected for different applications or may be used together.

After determining the learning granularity, the learning attributes should be determined. In ordinary relational databases, the attributes can be the fields explicitly stored or derived fields by mathematical or logical operation. On the contrary, the geometric features and spatial relations are not stored explicitly in spatial database, but hidden in the multi-layer graphic data. Spatial analysis and spatial operation must be performed to extract the attributes about shape

and spatial relation. For example, overlay analysis is needed to know which height zone an object falling into. This is a step of feature selection, which is a characteristic of spatial data mining.

Figure 1 is the flow diagram of inductive learning in spatial databases. Generally, learning samples are selected randomly from spatial database. When data storage is not very large, we can chose the whole data as learning data. After determining learning granularity and attributes, the learning data are organized to a tabular form as the input to C5.0 algorithm. C5.0 generates two kinds of outputs: decision tree and production rules. We chose production rules as the outputs because they are easy to understand and use.

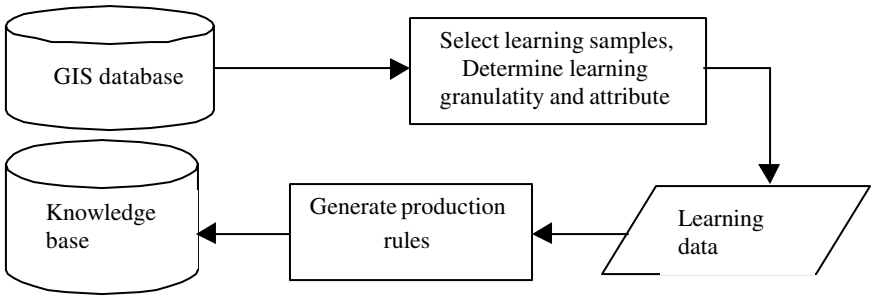

Figure 1. Flow diagram of inductive learning in spatial database

3 REMOTE SENSING IMAGE CLASSIFICATION BASED ON INDUCTIVE LEARNING

In the field of remote sensing, Bayes classification (or maximum likelihood classification) is most widely used. It can obtain minimum classification error under the assumption that the spectral data of each class is normally distributed. Generally, there is much spectral confusion between classes. That is, same class with different spectral and different class with the same spectral. The Bayes method itself can not solve the problem of spectral confusion. And because of the requirement of statistical distribution, the auxiliary data can not be incorporated in Bayes classification.

For most multi-spectral remote sensing data, Bayes method classifies the coarse classes correctly, such as water, residential area, green patches, etc. But usually more detailed classification is required in land use classification in China. For example, water should be subdivided into river, lake, reservoir and pond; green patch should be subdivided into vegetable field, garden, forest etc. These involve much spectral confusion. In order to subdivide water, shape information and spatial association knowledge should be used. In order to subdivide green patches, spatial distribution and also the slight difference should be used. In the following experiment, two kinds of knowledge are discovered from land use and elevation data, which are applied to subdivide water and green patches respectively.

Pixel granularity is adopted for learning knowledge to subdivide green patches. We propose an approach to combine inductive learning with Bayes classification method, which selects class probability of Bayes classification as learning attributes. Firstly, the image are classified by Bayes method, the probabilities of each pixel to every classes are retained. Then inductive learning is conducted taking probability values, location and elevation as the learning attributes. Since the probability is derived from the spectral information of a pixel and the statistical information of a class, learning with probability values makes use of the two kinds of information simultaneously. Comparative experiments show that using probability values generates more accurate learning results than using the pixel values directly. It indicates that this approach of combining inductive learning and Bayes method is effective.

Polygon granularity is adopted to subdivide waters. Knowledge about general geometric features and spatial distribution patterns are discovered from polygons of different waters. Before using the knowledge, the remote sensing image is classified first, the water areas in the classification image are converted form pixels to polygon by raster to vector conversion and then the location and shape features of these polygons are calculated. Finally, the polygons are subdivided into river, lake, reservoir and pond by deductive reasoning based on the knowledge. Here the combination of inductive learning and Bayes classification is in a loose manner.

Figure 2 shows the diagram of remote sensing image classification with inductive learning. GIS data are used in training area selection for Bayes classification, generating learning data of two granularities, test area selection for classification accuracy evaluation. And also the GCPs for image rectification are chosen from GIS data. Therefor, GIS plays important roles in remote sensing image classification from the beginning to the end.

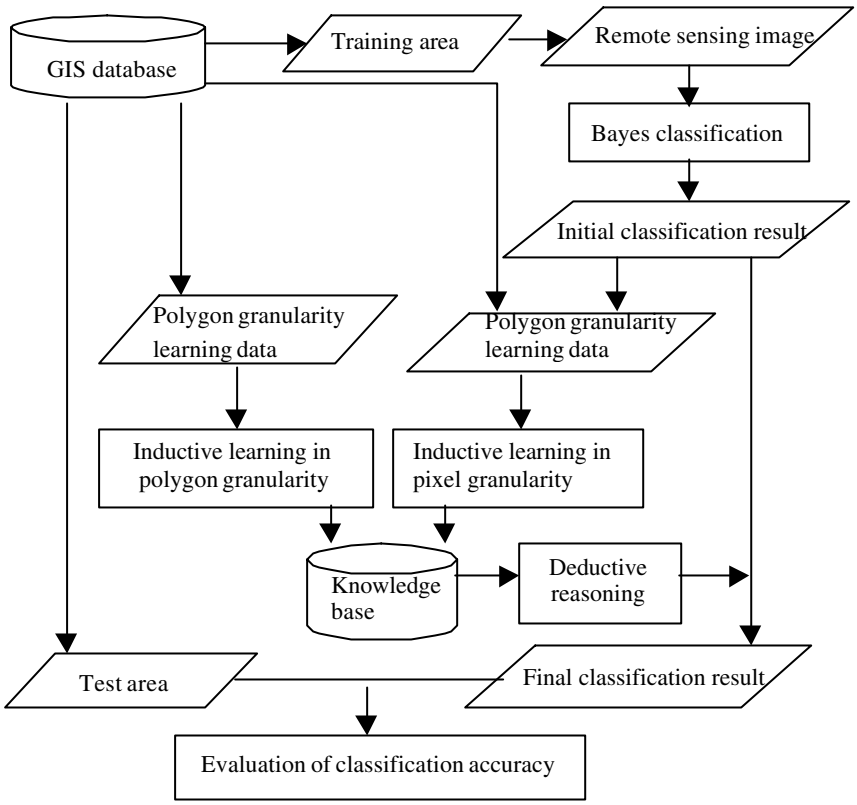

Figure 2. Flow diagram of remote sensing image classification with inductive learning

The knowledge discovered by C5.0 algorithm is a group of classification rules and a default class, and with each rule, there is a confidence value (between 0 and 1). As shown in figure 2, the final classification results are obtained by postprocessing of the initial classification results by deductive reasoning. The attributes for deductive learning are the same as that in inductive learning except the class label attribute. The following strategies are adopted in deductive reasoning: (1) If only one rule is activated, which means the attribute values match the conditions of this rule, let the final class be the same as this rule; (2) If several rules are activated, let the final class be the same as the rule with the maximum confidence; (3) If several rules are activated and the confidence values are the same, then let the final class be the same as the rule with the maximum coverage of learning samples; (4) If no rule is activated, , then let the final class be the default class.

The way of utilizing GIS information in data mining based image classification is quite different from the conventional way. Conventionally, GIS data are used directly in pre- or post- processing of image classification. In the data mining based image classification scheme, the knowledge, which was mined from the data, is used. Generally, knowledge is more generalized, condensed, reliable and easy to understand than data. And a group of rules can represent very complex non-linear knowledge. Therefore, utilizing knowledge is likely to be more beneficial to improve remote sensing image classification than utilizing GIS data directly.

4 LAND USE CLASSIFICATION EXPERIMENT

In order to verify the feasibility and effectiveness of the data mining based image classification, a land use classification experiment is performed in the Beijing area using SPOT multi-spectral image and 1: 100,000 land use database. The original image is 2412 by 2399 pixels and three bands, which was obtained in 1996. The land use database was built before 1996, which has land use, contour, road and annotation layers. The original image is stretched and rectified to the GIS data. The image is 2834 by 2824 pixels after rectification (See Fig.3), which is used as the source image for classification. We use ArcView 3.0a, ENVI 3.0 and See5 1.10, which is developed based on C5.0 algorithm by Rulequest Cooperation. And also we developed several programs for data processing and format conversion using Microsoft C^{++}5.0.

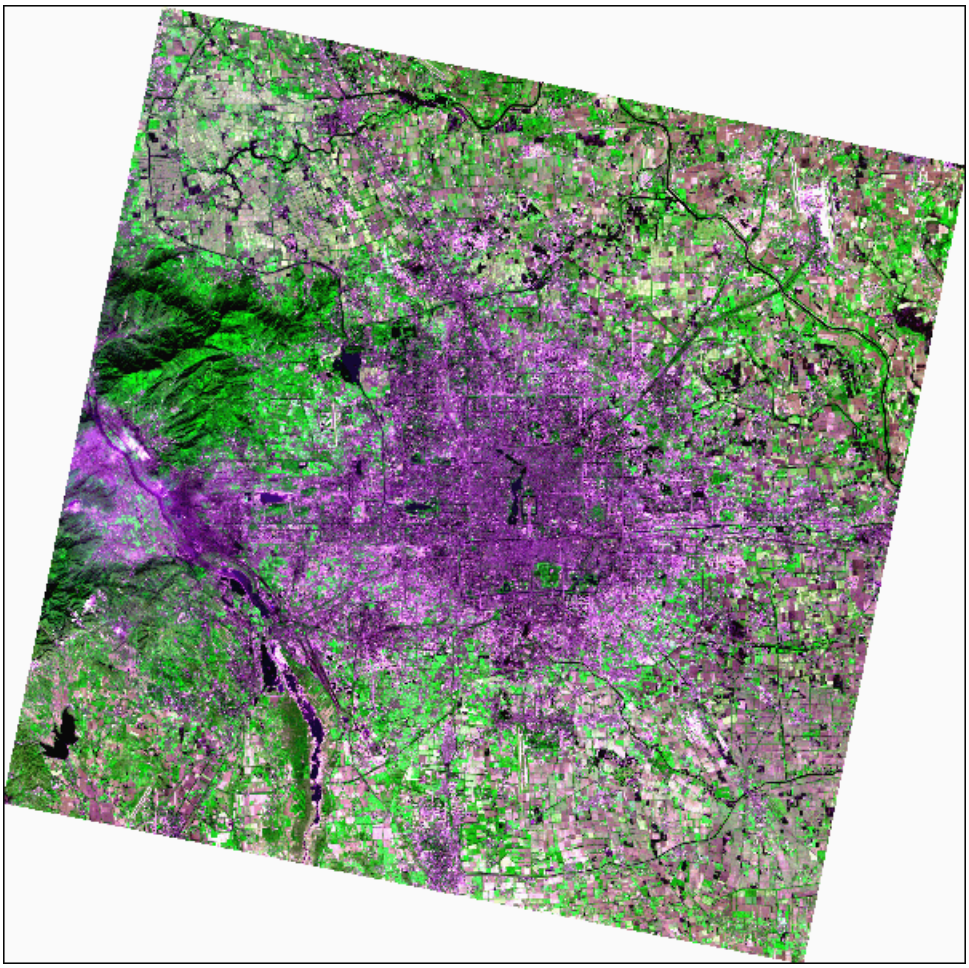

Figure 3. SPOT multi-spectral image for land use classification (resampled)

For the sake of comparison, only the Bayes method is applied to classify the image at first. The rectified image is overlaid with land use data layers, and the training and test areas are interactively selected. And then the image is classified into 8 classes, such as water, paddy, irrigated field, dry land, vegetable field, garden, forest and residential area. As shown in the confusion matrix (table 1), the overall accuracy is 77.6199%. Water, paddy, irrigated field, residential area and vegetable field are classified with high accuracy. The vegetable field is easlydistinguished from other green patches because it is lighter than the others are. Dry land, garden, forest are confused seriously and the accuracy is 65.58%, 48.913% and 59.754% respectively. And some forest shadows are misclassified as waters.

Classified	Real class								
	water	paddy	irrigated field	dry land	vegetable field	garden	forest	residential area	Sum
water	3.900	0.003	0.020	0.013	0.002	0.021	2.303	0.535	6.797
paddy	0.004	8.496	0.087	0.151	0.141	0.140	0.103	0.712	9.835
irrigated field	0.003	0.016	10.423	0.026	0.012	0.076	0.013	0.623	11.192
dry land	0.063	0.48	0.172	1.709	0.361	2.226	2.292	1.080	8.384
vegetable field	0.001	0.087	0.002	0.114	3.974	0.634	0.435	0.219	5.465
garden	0.010	0.009	0.002	0.325	0.263	4.422	4.571	0.065	9.666
forest	0.214	0.006	0.000	0.271	0.045	1.354	15.671	0.642	18.202
residential area	0.132	0.039	0.127	0.080	0.049	0.168	0.839	29.024	30.459
Sum	4.328	9.135	10.834	2.689	4.846	9.041	26.227	32.901	100
Accuracy (%)	90.113	93.010	96.204	63.580	81.994	48.913	59.754	88.217	
Overall accuracy = 77.6199% Kappa coefficient = 0.7474									

Table 1. Confusion matrix of Bayes classification

As stated in section 3, inductive learning is mainly used in two aspects to improve the Bayes method in land use classification, one is to discover rules to subdivide waters in polygon granularity, the other in to discover rules to reclassify dry land, garden and forest in pixel granularity. The land use layer (polygon) and contour layer (line) are selected for these purposes. Because there are few contours and elevation points, it is difficult to interpolate a DEM accurately, instead, the contours are converted to height zones, such as <50m, 50-100m, 100-200m and >200m, which are represented by polygons.

In the learning to subdivide waters, several attributes of the polygons in land use layer are selected or calculated as condition attributes, such as area, location of the center, compactness (perimeter^2/(4π*area)), height zone, etc. The classes are river (code 71), lake (72), reservoir (73), pond (74) and forest shadow (99). 604 water polygons are learned, 10 rules are discovered (See table 2). The are only 1.2% samples are misclassified in the learning, thus the learning accuracy is 98.8%. These rules reveal the spatial distribution patterns and general shape features, etc. For example, rule 1 states "If compactness of a water polygon is greater than 7.190882, and locates in the height zone <50m, then it is a river". Here the compactness measure plays a key role to identify river from other waters. Rule 2 identifies lakes by location and compactness, rule 9 and rule 10 distinguish forest shadows from waters by height, and so on.

```
Rule 1: (cover 19)
        compactness > 7.190882
        heiqht = lt50
        ->  class 71  [0.952]

Rule 2: (cover 5)
        Xcoord > 453423.5
        Xcoord <= 455898.7
        Ycoord > 4414676
        Ycoord <= 4428958
        compactness > 2.409397
        compactness <= 7.190882
        ->  class 72  [0.857]

Rule 3: (cover 33)
        Xcoord <= 455898.7
        Ycoord > 4414676
        Ycoord <= 4428958
        compactness <= 7.190882
        heiqht = lt50
        ->  class 72  [0.771]

Rule 4: (cover 4)
        area > 500000
        heiqht = 50_100
        ->  class 73  [0.667]

Rule 5: (cover 144)
        Ycoord <= 4414676
        compactness <= 7.190882
        heiqht = lt50
        ->  class 74  [0.993]

Rule 6: (cover 213)
        Ycoord > 4428958
        compactness <= 7.190882
        heiqht = lt50
        ->  class 74  [0.986]

Rule 7: (cover 281)
        Xcoord > 451894.7
        compactness <= 7.190882
        ->  class 74  [0.975]

Rule 8: (cover 38)
        area <= 500000
        heiqht = 50_100
        ->  class 74  [0.950]

Rule 9: (cover 85)
        heiqht = gt200
        ->  class 99  [0.989]

Rule 10: (cover 7)
        heiqht = 100_200
        ->  class 99  [0.778]

Default class: 74

        ----------------------
        Evaluation (604 cases):

        Errors      7( 1.2%)
```

Table 2. Rules discovered by inductive learning to subdivide waters

In the learning to reclassify dry land, garden and forest, the condition attributes are image coordinates, heights and the probability values to the three classes that produced by Bayes classification. One percent (2909) samples are selected randomly from the vast amount of pixels. 63 rules are discovered and the learning accuracy is 97.9%. The test accuracy is 94.4%, which was evaluated by another 1% randomly selected samples. These rules are omitted here because of paper size limitation.

After inductive learning, the Bayes classified image is reclassified by deductive reasoning based on the discovered rules. Because Bayes method can not subdivide waters, only the rules to identify forest shadows from waters are used in order to compare the result with Bayes classification. The final class is determined by the maximum confidence principle. And the final classification result in shown in figure 4. Accuracy evaluation is accomplished using the same test areas as that in Bayes classification. The confusion matrix is shown table 3. The overall accuracy of the final result is 88.8751%. The accuracy of dry land, garden and forest is 69.811%, 78.561% and 91.81% respectively. Comparing the final result with the result produced only by Bayes classification, the overall accuracy increased 11.2552 percent and the accuracy of dry land, garden and forest increased 6.231%, 29.648% and 32.056% respectively.

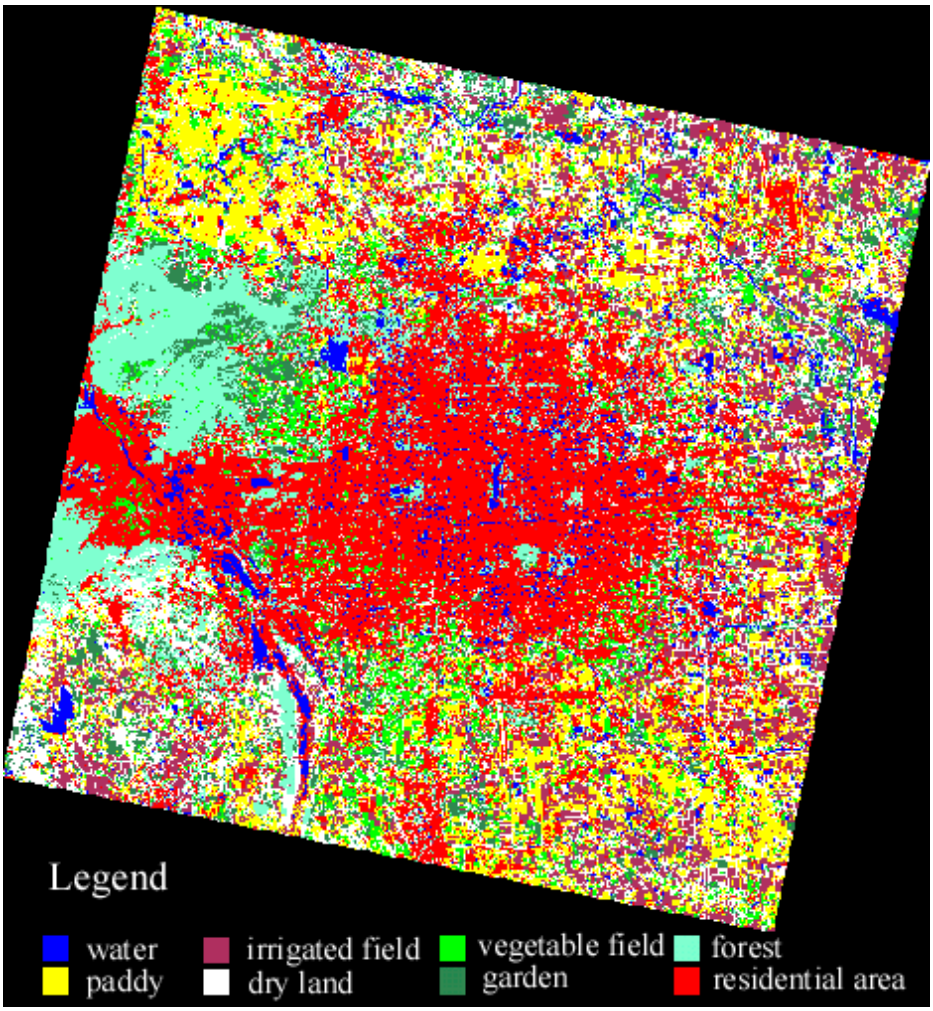

Figure 4. Classification result by combining Bayes method with inductive learning (resampled)

Classified	Real class								
	water	paddy	irrigated field	dry land	vegetable field	garden	forest	residential area	Sum
water	3.900	0.003	0.020	0.012	0.002	0.019	0.139	0.535	4.631
paddy	0.004	8.496	0.087	0.151	0.141	0.14	0.103	0.712	9.835
irrigated field	0.003	0.016	10.423	0.026	0.012	0.076	0.013	0.623	11.192
dry land	0.063	0.480	0.172	1.877	0.361	0.205	0.149	1.080	4.386
vegetable field	0.001	0.087	0.002	0.114	3.974	0.634	0.435	0.219	5.465
garden	0.009	0.009	0.002	0.210	0.263	7.102	0.470	0.065	8.131
forest	0.215	0.006	0.000	0.218	0.045	0.696	24.079	0.642	25.899
residential area	0.132	0.039	0.127	0.080	0.049	0.168	0.839	29.024	30.46
Sum	4.328	9.135	10.834	2.689	4.846	9.041	26.227	32.901	100
Accuracy (%)	90.113	93.01	96.204	69.811	81.994	78.561	91.81	88.217	
Overall accuracy = 88.8751% Kappa coefficient = 0.8719									

Table 3. Confusion matrix of Bayes classification combined with inductive learning

5 CONCLUSION

The experiment result of land use classification shows that the overall accuracy increased more than 11% and the accuracy of some classes, such as garden and forest, increased about 30 percent. This indicates that spatial data mining

techniques are very helpful to improve the traditional Bayes classification method and the proposed approaches of the implementation of inductive learning in spatial databases are feasible and effective. In inductive learning can resolve the problem of spectral confusion to a great extent. Combining Bayes method with inductive learning not only improves classification accuracy greatly, but also extends the classification by subdivide some classes with the discovered knowledge.

The intelligent integration of GIS and remote sensing is a difficult problem. An encouraging solution to the problem is mining knowledge from spatial and utilizing the knowledge in image interpretation for spatial data updating. The implementation of inductive learning in spatial databases and the combination with traditional classification methods are theoretically and practically valuable. The applications of inductive learning to other image data sources, such as TM, SAR etc., and the applications of the other data mining methods in remote sensing image classification, are the future directions of our further study.

REFERENCES

Eklund P.W., Kirkby S.D, A. Salim 1998. Data mining and soil salinity analysis. Int. J. Geographical Information Sciences, Vol. 12, No 3, pp247-268

Hong Jiarong 1997. Inductive learning – algorithm, theory and application, Science Press, Beijing, Sept.

Huang Xueqiao and John R. Jensen 1997. A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIS Data. Photogrammetric Engineering & Remote Sensing, Vol.63, No10, pp1185-1194

Li Deren, Cheng Tao 1994. KDG: Knowledge Discovery from GIS - Propositions on the Use of KDD in an Intelligent GIS. In Proc. ACTES, The Canadian Conf. on GIS

Li Deren, Di Kaichang, and Li Deyi 1997. A Framework of spatial data mining and knowledge discovery. In Proc. Int. Workshop on Image Analysis and Information Fusion(IAIF'97), Adelaide, Australia, Nov.

Li Deyi 1992. Inductive learning: knowledge discovery from database. In: Proc. of 10th National Conf. on Database, Shenyang, China, Sept.

Quinlan J.R 1993. C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA

Zhang Jixian et al. 1995. Methods and key techniques of GIS database updating based on remote sensing image source. In: Proc. 1st Annual Conf. of China Association on GIS, Beijing