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

DI Kaichang
LI Deren
LI Deyi

KEY WORDS data mining; knowledge discovery; image classification; inductive learning; learning granularity

ABSTRACT Data mining techniques are used to discover knowledge from GIS database in order to improve remote sensing image classification. Two learning granularities are proposed for inductive learning from spatial data, one is spatial object granularity, the other is pixel granularity. 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 by 11% and the accuracy of some classes, such as garden and forest, increased by about 30%. The results indicate that inductive learning can resolve spectral confusion to a great extent. Combining Bayes method with inductive learning not only improves classification accuracy greatly, but also extends the classification by subdividing some classes with the discovered knowledge.

1 Introduction

It has long been 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 fulfilled in a statistical or interactive manner, so it is difficult to use the auxiliary data automatically and intelligently.

Furthermore, 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 image classification expert system was very difficult because of the “knowledge acquisition bottleneck”.

Spatial data mining and knowledge discovery (SDMKD), is the extraction of implicit, interesting spatial or non-spatial patterns and general characteristics. We proposed a theoretical and technical framework of spatial data mining and knowledge discovery (Li et al., 1997). And spatial data min-
Remote Sensing Image Classification with GIS Data Based on Spatial Data Mining Techniques

Remote sensing image classification with GIS data is supposed to be used in two aspects: one is intelligent analysis of GIS data, the other is to support knowledge-driven interpretation and analysis of remote sensing images. SDMDK is a new method of knowledge acquisition for remote sensing image classification. Several researchers have done work in this field. Eklund et al. (1998) extracted knowledge from TM images and geographic data in soil salinity analysis using inductive learning algorithm C4.5. Huang et al. (1997) extracted knowledge from GIS data and SPOT multispectral image in wetland classification using C4.5 too. 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, still need to be further studied.

2 Implementing inductive learning 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 the most important one. And there are many inductive learning algorithms which mainly come from the field of machine learning, for example, AQ 11 and AQ 15 by Michalski, AE1 and AE9 by Hong Jiarong, 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 the most famous and influential. C5.0 is an improvement over ID3 and C4.5 and it runs much faster in very large databases. Therefore, when studying the implementation of C5.0, as many other inductive learning algorithms, we 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. Because spatial data structure is more complex than the tables in ordinary relational database, learning in spatial database is more difficult than learning in relational database for selecting the tuple and attributes of training data.

We regard learning tuple selection as a problem for 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 areal, 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 of spatial object granularity can discover knowledge about location, shape, spatial relation, etc. The discovered knowledge is generalized and can be used in intelligent spatial data analysis and in remote sensing image classification as well. When the discovered rules are applied to image classification, the image must be clustered or pre-classified to areal or linear features before the rules are used. Learning of 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 of pixel granularity can not utilize shape information and it is difficult to utilize spatial association information. Learning of spatial object granularity can not utilize the detailed 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 mean or sample value. These two kinds of granularities should be chosen for different applications or should be used together.

After determining the learning granularity, the learning attributes should be determined. Usually, 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. This is a step of feature selection, which is a characteristic of spatial data mining. Fig. 1 is the flow diagram of inductive learning in spatial databases.
mining 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 shows the flow diagram of inductive learning in the 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. For most multi-spectral remotely sensed data, by Bayes method the coarse classes, such as water, residential area, green patches, etc. can be classified correctly, 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. Those involve much spectral confusion. The Bayes method itself is not capable of solving this problem. In order to subdivide water, shape information and spatial association knowledge should be used. In order to subdivide green patches, spatial distribution and the slight difference between them should be used as well.

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 by taking probability values, location and elevation as the learning attributes. Since the probability is derived from the spectral information of 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 of the three bands.

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 by Bayes method, the water areas in the image being classified are converted from 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.

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

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 Fig. 2, the final classification results are obtained by postprocessing of the initial classification results by deductive reasoning. The maximum confidence principle is adopted in deductive reasoning when several rules are activated simultaneously.

4 A land use classification experiment

In order to verify the feasibility and effectiveness of the data mining based on image classification, a land use classification experiment was performed in the Beijing area using SPOT multi-spectral image
Remote Sensing Image Classification with GIS Data Based on Spatial Data Mining Techniques

and 1:100,000 land use database. The image was obtained in 1996. The land use database was built before 1996, which has land use, contour, road and annotation layers. First, the original image was rectified to the GIS data. The image is of 2,834 by 2,824 pixels after rectification, which is used as the source image for classification. In the experiment we used ArcView 3.0a, ENVI 3.0 and See5 1.10, which is developed on the basis of C5.0 algorithm by Rulequest Cooperation. And we developed several programs for data processing and format conversion using Microsoft C++ 5.0 as well.

![Flow diagram of remote sensing image classification by inductive learning](image)

For the sake of comparison, only the Bayes method was applied to classify the image at first. 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 77.619%. Water, paddy, irrigated field, residential area and vegetable field are classified with high accuracy. The vegetable field is distinguished from other green patches because it is brighter than the others. Dry land, garden, forest are confused seriously and the accuracy is 63.58%, 48.913% and 59.754% respectively. Some forest shadows are misclassified as waters.

| Classified objects | 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/%

| water | paddy | irrigated field | dry land | vegetable field | garden | forest | residential area |
|-------|-------|-----------------|---------|-----------------|--------|--------|------------------|
| 90.113| 93.010| 96.204          | 63.580  | 81.994          | 48.913 | 59.754 | 88.217           |

Overall accuracy = 77.619% Kappa coefficient = 0.7474
As stated in section 3, inductive learning is mainly used to improve the Bayes method in land use classification from two aspects, one is to discover rules to subdivide waters in polygon granularity, the other is 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 $<50$ m, $50 \sim 100$ m, $100 \sim 200$ m and $>200$ m, which are represented by polygons.

In the learning to subdivide waters, several attributes of the polygons in land use layer were selected or calculated as condition attributes, such as area, location of the center, compactness (perimeter $\cdot 2/(4 \cdot$ area)), height zone, etc. The classes are river (code 71), lake (72), reservoir (73), pond (74) and forest shadow (99). 604 water polygons were learned, 10 rules were discovered. Only 1.2% samples were 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 $<50$ m, then it is a river." Here the compactness measure plays a key role to identify river from other waters. These rules are not shown here because of paper size limitation.

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. 1% samples were 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% by another 1% randomly selected samples. These rules are also not shown here.

After inductive learning, the classified image by Bayes method is reclassified by deductive reasoning based on the discovered rules. Because Bayes method can not be used to subdivide waters, only the rules to identify forest shadows from water are used in order to compare the result with Bayes classification. The final class was determined by the maximum confidence principle. Accuracy evaluation was accomplished using the same test areas as that in Bayes classification. The confusion matrix is shown in Table 2. The overall accuracy of the final result is 88.875%. 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 has increased by 11.225% and the accuracy of dry land, garden and forest by 6.231%, 29.648% and 32.056% respectively.

Table 2  Confusion matrix of Bayes classification combined with inductive learning

| Classified objects | 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    | 0.009 | 0.009 | 0.002          | 0.114    | 3.974          | 0.634  | 0.435  | 0.219            | 5.465 |
| garden             | 0.215 | 0.006 | 0.000          | 0.218    | 0.045          | 0.696  | 0.470  | 0.642            | 25.899|
| forest             | 0.132 | 0.039 | 0.127          | 0.080    | 0.049          | 0.168  | 0.839  | 29.024           | 30.46 |
| residential area   | 4.328 | 9.135 | 10.834         | 2.689    | 4.846          | 9.041  | 26.227 | 32.901           | 100   |
| Sum                | 90.113| 93.01 | 96.204         | 69.811   | 81.994         | 78.561 | 91.81  | 88.217           |       |

Accuracy/%

Overall accuracy = 88.875 1%  
Kappa coefficient = 0.871 9
5 Conclusion

The experiment results of land use classification show that the overall accuracy has increased by more than 11% and the accuracy of some classes, such as garden and forest, increased by about 30%. That 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. The inductive learning can resolve spectral confusion to a great extent. The combination of Bayes method with inductive learning not only improves classification accuracy greatly, but also extends the classification by subdividing some classes with the discovered knowledge.

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 direction of further study.

References

1. Eklund P W, Kirkby S D, Salim A. Data Mining and Soil Salinity Analysis. Int. J. Geographical Information Sciences, 1998, 12(3): 247–268
2. Hong J R. Inductive Learning-AlGORITHM, Theory and Application. Beijing: Science Press, 1997 (in Chinese)
3. Huang X Q, Jensen J R. A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIS Data. Photogrammetric Engineering & Remote Sensing, 1997, 63(10): 1185–1194
4. Li D R, Cheng T. KDG: Knowledge Discovery from GIS-Propositions on the Use of KDD in an Intelligent GIS. The Canadian Conf. on GIS, 1994
5. Li D R, Di K C, Li D Y. A Framework of Spatial Data Mining and Knowledge Discovery. In: Proc. Int. Workshop on Image Analysis and Information Fusion (IAIF’97). Australia: Adelaide, 1997
6. Li D Y. Inductive Learning: Knowledge Discovery from Database. In: Proc. of 10th National Conf. on Database. Shenyang, 1992 (in Chinese)
7. Quinlan J R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993
8. Zhang J X. 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, 1985 (in Chinese)