CBERS-02 Remote Sensing Data Mining Using Decision Tree Algorithm

Xingping Wen, Guangdao Hu
Institute of Mathematic Geology and Remote Sensing Geology
China University of Geosciences
Wuhan, China
wfxyp2008@gmail.com

Xiaofeng Yang
College of Environmental Science and Engineering
Nanjing University of Information Science & Technology
Nanjing, China

Abstract

In recent years, decision tree algorithms have been successfully used for land cover classification from remote sensing data. In this paper, CART (classification and regression trees) and C5.0 decision tree algorithms were used to CBERS-02 remote sensing data. Firstly, the remote sensing data was transformed using the Principal Component Analysis (PCA) and multiple-band algorithm. Then, the training data was collected from the combining total 20 processed bands. Finally, the decision tree was constructed by CART and C5.0 algorithm respectively. Comparing two results, the most important variables are clearly $band_{3,4}$, $band_{1,4}$ and $band_{2,4}$. The depth of the CART tree is only two with the relative high accuracy. The classification outcome was calculated by CART tree. In order to validate the classification accuracy of CART tree, the Confusion Matrices was generated by the ground truth data collected using visual interpretation and the field survey and the kappa coefficient is 0.95.

1. Introduction

Remotely sensed image data is widely used in a range of oceanographic, terrestrial, and atmospheric applications, such as land cover mapping, environmental modeling and monitoring [1]. For more than a decade, pattern recognition methods applied to remotely sensed imagery classification have mainly been based on conventional statistical techniques, such as the maximum likelihood or minimum distance procedures, using a pixel-based approach. They are classified into two main categories known as the supervised and unsupervised approaches [2]. Unsupervised classification methods are less dependent on user interaction, but its accuracy is generally lower than that achieved by supervised methods. Supervised methods require the user to collect samples to 'train' or teach the classifier to determine the decision boundaries in feature space, however, the rendering rule can not be applied to the other data set. Furthermore, some methods frequently assume that the data follow Gaussian distribution, an assumption which may not be tenable in remote sensing images containing mixed pixels, so their general ability for resolving inter-class confusion is limited. As a result, in recent years, following advances in computer technology, alternative data mining strategies have been proposed, particularly the use of artificial neural networks, decision trees, methods derived from fuzzy set theory. Decision tree classifiers have been successfully used for land cover classification from remote sensing data [3-7]. They successively partition the input data into more and more homogeneous subsets by producing optimal rules which minimize the error rates in the branches of the tree [8]. In addition, they can gain a more comprehensive understanding of relationships between objects at different scales of observation or at different levels of detail [1]. Moreover, its rendering rule can apply to other data sets. The research described in this paper, CART and C5.0 algorithms for building decision trees are used in the CBERS-02 remote sensing data.

e-Forensics 2008, January 21-23, 2008, Adelaide, Australia.

2. Materials and methods

2.1. CBERS remote sensing data and study areas

The CBERS (China Brazil Earth Resources Satellite) was jointly developed by China and Brazil since 1988. There are three kind cameras on the CBERS-02. This paper used the CCD images acquired in October 10, 2004. The CCD Camera has a nadir spatial resolution of 19.5 meters and a swath width of 113Km. It has four spectral bands in the visible and near infrared range and one panchromatic band. Total 5 bands are available. This study used the image located at Guangzhou of China. Guangzhou is the capital city of Guangdong Province, situated in the Pearl River Delta in South China, which has twelve districts (development zones) and county-level cities. The study area covers the centre seven districts (fig. 1c) between latitude

23°02'N - 23°26'N and longitude113°08'E - 113°36'E. Before using data mining tools, the image was geo-referenced to a Transverse Mercator projection with an RMSE of 0.5 pixels.

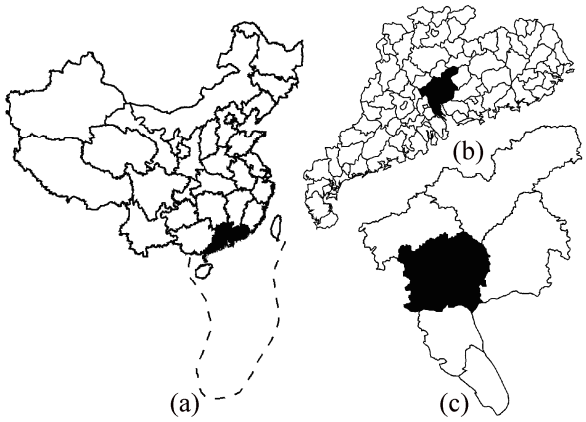

Figure 1. China provincial boundary map, the black area is Guangdong provinces (a). The county boundary map of Guangdong provinces, the black area is Guangzhou municipality (b). Guangzhou district boundary map adjusted in 2005, the black area is the study areas (c).

2.2. CBERS -02 remote sensing data processing

A total of 4363 point samples were randomly collected from the remote sensing data and each point sample was assigned to one class based on the visual interpretation and the field data, then these data were calculated using decision tree algorithm. The decision tree algorithm produces a tree of more variable shape when only using the origin 5 bands data, so some data processing are used. In this paper, the PCA and multiple-band algorithm were used. The multiple-band algorithm was defined by (1).

$$band_{i,j} = \frac{band_i - band_j}{band_i + band_j} \qquad (1)$$

$(i, j=1,2,3,4,5)$

Where: i, j refer to band number shown in table 1.

Table 1. The band number of CBERS-02 CCD image with the corresponding wavelength.

	Band number	Wavelength
VIS/NIR	1	0.45～0.52
	2	0.52～0.59
	3	0.63～0.69
	4	0.77～0.89
panchromatic	5	0.51～0.73

Every two band was computed by (1), so 10 new bands data were computed out. PCA transformation can output 5 new bands data. Adding the origin 5 bands, total 20 bands data were used in decision tree algorithm.

2.3. The decision tree algorithm

In this paper, CART and C5.0 algorithm were used. CART was suggested by Breiman et al. in 1984 [9]. The decision trees produced by CART are strictly binary, containing exactly two branches for each decision node. CART recursively partitions the records in the training data set into subsets of records with similar values for the target. The C5.0 algorithm is the successor of the C4.5 program by Quinlan [10]. Just as with CART, C5.0 algorithm recursively visits each decision node, selecting the optimal split, until no further splits are possible. However, C5.0 algorithm for measuring node homogeneity is quite different from the CART method [11]. A number of recent studies have reported the classification performance of CART and C5.0, with different success levels [12-15]. In this paper, their outcome trees were compared with each other.

3. Result

All sample data were split into two parts. 80 percent sample data were used as the training data and 20 percent sample data were used as across validation data. The depth and across validation accuracy of decision tree constructed by C5.0 and CART algorithm using the processed 20 bands data are shown in table 2.

Table 2. The depth and cross-validation accuracy of the decision tree constructed by C5.0 and CART algorithm.

Algorithm	Tree depth	Accuracy
C5.0	5	99.77%
CART	2	99.32%

Although the CART and C5.0 decision trees do not agree in the details, there is some agreement between them, the most important variables are clearly band $_{3,4}$, band $_{1,4}$ and band $_{2,4}$. Comparing two results, the CART tree is so concise with the relative high accuracy that it will have high generalization capability. Fig. 2 is the CART decision tree and fig.3 is the plot of band $_{1,4}$ versus band $_{3,4}$. As is shown in fig.2 and fig. 3, the band$_{3,4}$ known as NDVI (Normalized Difference Vegetation Index) [16] is the most important variable. It can distinguish the vegetation from the remote sensing data. Band$_{1,4}$ can recognize the water, and band$_{2,4}$ can discriminate the forest from the vegetation. In order to testify the classification accuracy of the

CART tree, the ground truth data was collected using visual interpretation and the field survey, then Confusion Matrices was generated and the kappa coefficient [17] is 0.95. At last, the classification result was outputted. Fig. 4 is the classification image of land use in Guangzhou of China.

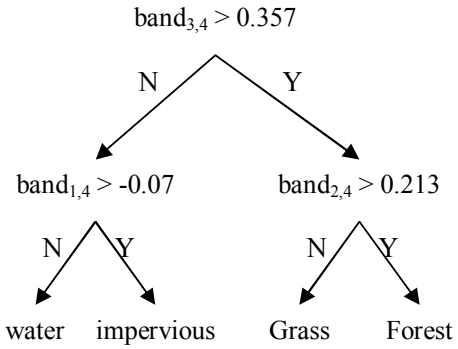

Figure 2. The decision tree constructed by CART algorithm.

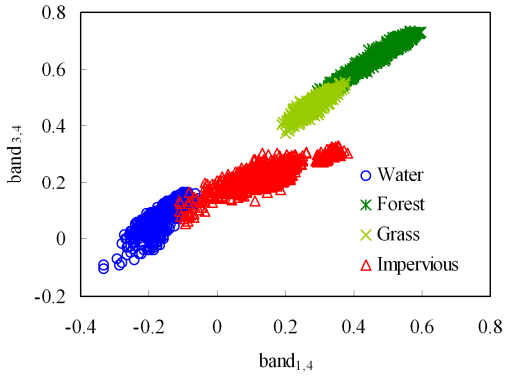

Figure 3. The plot of band $_{1,4}$ versus band $_{3,4}$.

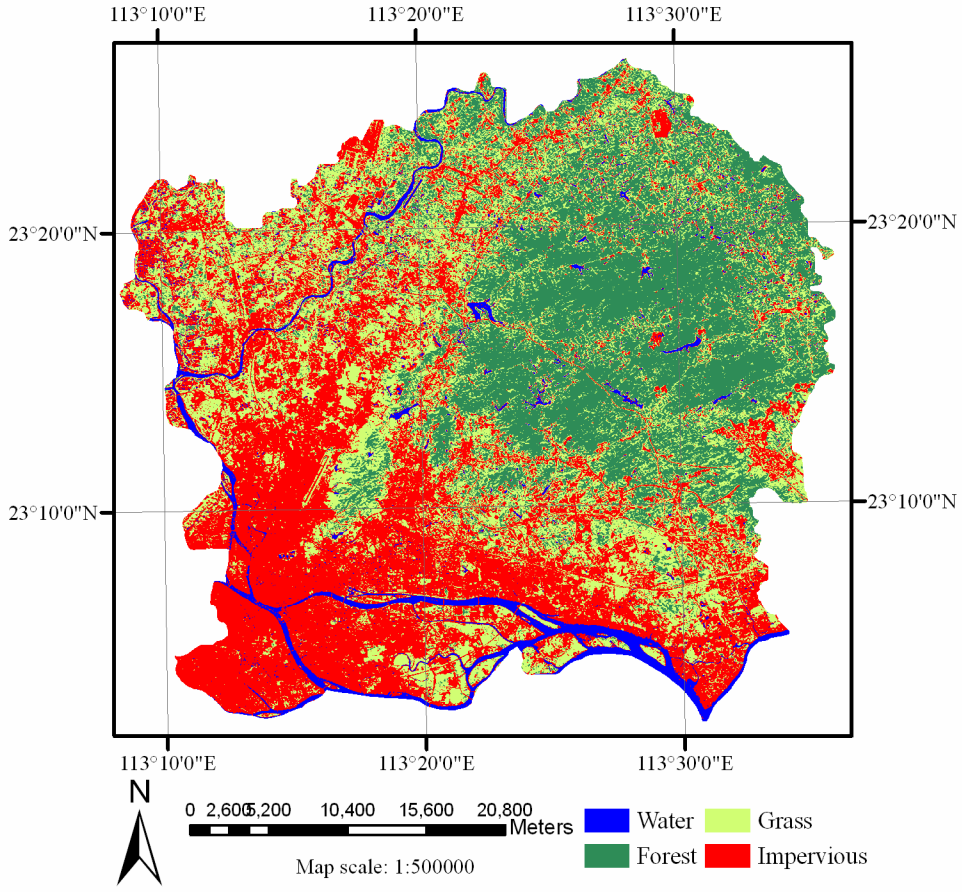

Figure 4. The classification image of land use in Guangzhou in China.
Map Projection: Transverse Mercator, False Easting: 500 kilometer, Central Meridian: 111°.

4. Acknowledgment

The authors would like to thank China Center for Resource Satellite Data and Applications (CRESDA) for providing the CBERS-02 remote sensing image. This study was supported by National Land and Resources Survey Foundation of China (No. 2003024002).

5. References

[1] B. Tso and P. M. Mather, *Classification Methods for Remotely Sensed Data*. London and New York: Taylor & Francis New York, 2001.

[2] J. A. Richards, *Remote Sensing Digital Image Analysis*. Berlin: Springer-Verlag, 1999.

[3] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data," *Remote Sensing of Environment,* vol. 97, pp. 322-336, 2005.

[4] C.-C. Yang, S. O. Prasher, P. Enright, C. Madramootoo, M. Burgess, P. K. Goel, and I. Callum, "Application of decision tree technology for image classification using remote sensing data," *Agricultural Systems,* vol. 76, pp. 1101-1117, 2003.

[5] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sensing of Environment,* vol. 61, pp. 399-409, 1997.

[6] K. C. Lee and S. J. Park, "A knowledge-based fuzzy decision tree classifier for time series modeling," *Fuzzy Sets and Systems,* vol. 33, pp. 1-18, 1989.

[7] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sensing of Environment,* vol. 86, pp. 554-565, 2003.

[8] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 21, pp. 660-673, 1991.

[9] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC Press, 1984.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[11] D. T. Larose, *Discovering Knowledge in Data an Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2005.

[12] J. Im and J. R. Jensen, "A change detection model based on neighborhood correlation image analysis and decision tree classification," *Remote Sensing of Environment,* vol. 99, pp. 326-340, 2005.

[13] P. K. Goel, S. O. Prasher, R. M. Patel, J. A. Landry, R. B. Bonnell, and A. A. Viau, "Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn," *Computers and Electronics in Agriculture,* vol. 39, pp. 67-93, 2003.

[14] E. C. Brown de Colstoun and C. L. Walthall, "Improving global scale land cover classifications with multi-directional POLDER data and a decision tree classifier," *Remote Sensing of Environment,* vol. 100, pp. 474-485, 2006.

[15] T. Waheed, R. B. Bonnell, S. O. Prasher, and E. Paulet, "Measuring performance in precision agriculture: CART--A decision tree approach," *Agricultural Water Management,* vol. 84, pp. 173-185, 2006.

[16] C. J. Tucker, "Red and Photographic Infrared Linear Combinations for Monitoring Vegetation," *Remote Sensing of Environment,* vol. 8, pp. 127-150, 1979.

[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, pp. 37-46, 1960.

Corresponding author: Xingping Wen

The mailing address:
Institute of Mathematic Geology and Remote Sensing Geology; China University of Geosciences;
Wuhan; The People's Republic of China;

Post code: 430074

Corresponding author email: wfxyp2008@gmail.com

Telephone Number: (086)-027-63350971