REALIZATION OF DATA MINING MODEL FOR EXPERT CLASSIFICATION USING MULTI-SCALE SPATIAL DATA

Shuang Zhang, Xuehua Liu*

Department of Environmental Science and Engineering, Tsinghua University, Beijing 100084, China - Zhangshuang03@mails.tsinghua.edu.cn, Xuehua-hjx@mail.tsinghua.edu.cn

KEY WORDS: Decision Tree, Knowledge Classification, Remote Sensing, GIS, Landscape

ABSTRACT:

Data mining models show great efficiency on acquiring knowledge for expert system classification. This study aimed at mining knowledge contained in landscape from multi-scale spatial data using decision tree learning model and evaluating the classification quality influenced by different scales of spatial data. Firstly, spatial data containing remote sensing images of different spatial and spectrum resolutions, digital elevation models and geographical information data with different scales were combined together to make up a spatial data infrastructure. Secondly, field samples data acquired by GPS were taken as the reference and the related spatial data was extracted. Thirdly the expert rules were developed by C5.0 decision tree models and then the rule base was used in a knowledge classifier. Finally we measured the accuracy influence of the data and data sets with different scales. The results showed: (1) The potential knowledge and rules could be detected using this data mining model with enough field samples. (2) The information provided by multi-level spatial data would influence the decision tree learning. Data set with a scale of 20m would offer most effective information. (3) After selecting effective data and scaling, we got an acceptable accuracy of 80.7% using the decision tree data mining and expert classification.

1. INTRODUCTION

As being widely commented, knowledge acquirement and rule base construction is one of the most important processes in expert system classification (Joseph and Gary, 2002). However, it is such a difficult work that often limits the application of expert system classification (Geoffery, 1996). Nowadays data mining models are designed and used to acquire knowledge form the available data. According to these models, knowledge and rules could be discovered from the data sets automatically (Mehmed, 2002; Shi, 2002). Also data mining models could be used to acquire spatial rules and knowledge for remote sensing (RS) expert system classification. Through these mining models we could get higher accuracy classification results (Di, 2001).

However, the accuracy of RS classification could be influenced by several factors. One of the important factors is the scale or resolution of the spatial data. As for knowledge-based RS classification, different types of spatial data will be used, such as RS images, DEM, GIS vectors and so on. Because of the different data sources, sometimes there could be various resolutions or spatial scales among the data sets. So once a model was used for classification according to a data set, it is necessary to specify how the accuracy would change when we add other assistant data with a different spatial scale.

In this study we developed a knowledge-based classifier and carried out the whole classification process containing data mining procedure. Based on a set of data with different spatial scales, we classified the landscape of the Foping Nature Reserve in the southern slope of the Qinling Mountains, China. During the data mining we used the C5.0 decision tree model and all the knowledge extracted from the data source was regulated into the rule base for the further knowledge classification. Also we carried out several experiments using the same model but different data sources by adding, removing some data with different spatial scales or change the scale of the whole data set. According to this study, we would find the influence of the data set with different spatial scales to data mining model and carry out the means to reduce this influence.

2. METHODOLOGY

2.1 Study Site

The study site was located in the southern slope of the Qinling Mountains, China, where there is complex topography and diverse vegetation types. There are a series of nature reserves (NRs) in the Qinling Mountains that form a conservation network for large mammals such as the giant panda (*Ailuropoda melanoleuca*) and golden takin (*Budorcas taxicolor*). However, before these NRs were founded, most forests in the Qinling Mountains were influenced by human activities such as woodcutting, road construction, farming, and gathering of medicinal plants. The landscape became intricate after the disturbance and the vegetation restoration.

We focused our study site in Foping NR (107°40’ to 107°55’ E and 33°31’ to 33°44’ N), which lies in the southern of Shannxi Province. In order to measure the effect of the classification conveniently, we select an area of 10*10 km as our study sample, which cover an altitude range of 1178 to 2420 m (Figure 1). In this area there are different vegetation types as the topography influences. Also there are some remaining farm and settlement near the river basin. We aimed to classify the

* Corresponding author: Liu Xuehua, Tel.: +86-10-62785610 ext 14; Fax: +86-10-62785687.

landscape distribution in this area using the knowledge based classification model.

2.2 Data Source

Since the year 1999 we have finished several field surveys (1999, 2003, and2004) and collected over 500 sample points. Also we got field data from Third National Giant Panda Survey containing about 1500 sample points. In this study we selected 750 points in the study area as the reference data and all the sample points took the error of no more than 10m (Figure 1). According to the field survey, 9 types of landscape were specified: conifer forest (CF), mixed broadleaf and conifer forest (MBC), broadleaf forest (BF), bamboo (BAM), shrub/grass/herb (SGH), farmlands (FAR) settlements (SET), water (WA), rock and bare land (RB). To check the result of the classification, in this study we use 375 points in classification and 375 ones for accuracy analysis.

An ETM+ satellite image acquired on May 22nd, 2001 containing 7 bands (1, 2, 3, 4, 5, 7, and 8) was used as the main classification data in this study (Figure 1). Also we collected other spatial information such as NDVI distribution DEM, slope and aspect data, distance to the roads and rivers distribution in the reserve (Table 1). All the sample points and spatial data were integrated into the ArcGIS environment (UTM projection, WGS84 datum).

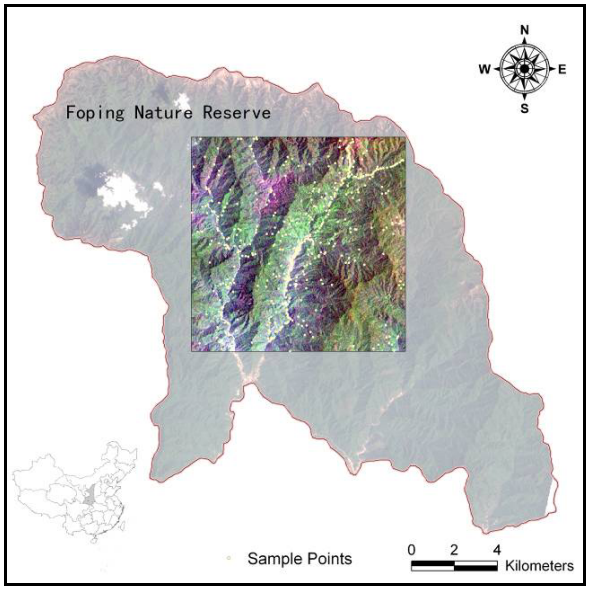

Figure 1. The study area of our research. The Foping NR lies in the southern slope of the Qinling Mountains founded for giant panda conservation. We select an area of 10*10 km in the middle of this reserve as our study sample. The main data source is an ETM+ image acquired on May 22nd, 2001 (this map used its combination of band 3, 2, and 1). We select 750 sample points as the reference data.

2.3 C5.0 Decision Tree

The decision tree (DT) learning model was more and more often used in RS classification these years (Huang and John, 1997; Eric et. al., 2003; Liu et. al., 2005). The advantages that DTs offer include an ability to handle data measured on different scales, no assumptions concerning the frequency distributions of the data in each of the classes, flexibility, and ability to handle non-linear relationships between features and classes (Friedl and Brodley, 1997). DTs could be trained quickly, and are rapid in execution. Besides, according to the DT learning model, knowledge could be realized with high accuracy (Shi, 2002).

Data	Description	Scale/Precision
Sample points	Gathered from the study since 2003	10m
RS image	ETM+ (band 1, 2, 3, 4, 5, 7, and 8) Acquired on May 22nd, 2001	28.5m (band 1, 2, 3, 4, 5, and 7); 14.25m (band 8)
NDVI	Derived from the RS image (TM4-TM3)/(TM4+TM3)	28.5m
DEM	Digitized form the paper map (1:50000)	25m
Aspect	Calculated from DEM	25m
Slope	Calculated from DEM	25m
GIS data	Distance to roads and rivers distribution in the reserve	10m

Table 1. The data source used in this research. We collected sample points as reference information. An ETM+ image and the calculated NDVI were used as the main data set. Also we acquired the DEM, aspect and slope data, road and river data to construct a classification data set.

DT uses a multi-stage or sequential approach to the problem of label assignment. Sets of decision sequences form the branches of the DT, with tests being applied at the nodes. The leaves (or branch termini) represent class labels (Figure 2). In this study a See5 DT model based on C5.0 algorithm was used to acquire the knowledge from the data sets. The C5.0 algorithm is a kind of univariate DT improved from the ID3 algorithm, which selects the branch feature according to the decrease rate of the information uncertainty calculated by equation 1 (Quinlan, 1993):

$$H(X/a) = -\sum_{j}\sum_{i} p(a=a_j)p(C_i/a=a_j)\log p(C_i/a=a_j) \quad (1)$$

Where a = the value of one feature
C = the class label
H (X/a) = information uncertainty of feature a

The feature with the minimal H(X/a) will be selected as the branch one.

2.4 Rule base knowledge classification

The rule-type of knowledge could be used for classification more effectively than the tree-type. So the classification tree would be converted into rules after finished DT learning in See5. In this study the knowledge engineer in ERDAS 8.7 environment was used to build the knowledge base. In this engineer, all the classes will be treated as hypothesises and will be concluded from several conditions of variables according to the rules we got from the DT (Figure. 3).

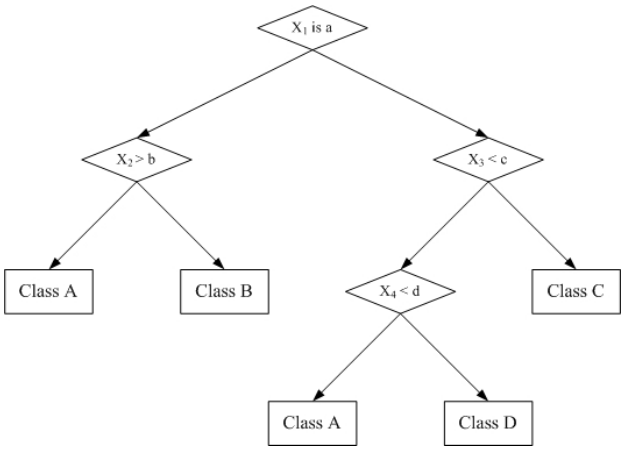

Figure 2. A univariate classification tree with four features and four classes. The xi are feature values; a, b, c, d are the thresholds and A, B, C, D are class labels. All the tests will be carried out at the nodes.

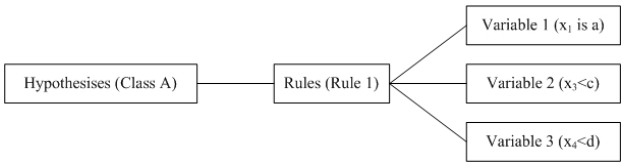

Figure 3. A rule-type displayed in ERDAS Knowledge Engineer from the tree in Figure 3. According to the conditions of variable x1, x3, and x4, a rule of the class A was founded.

2.5 Scale affection analysis

In order to check whether the data source with different scales would affect the classification accuracy, we repeated the DT learning and knowledge engineer classification for several times by select different data source in our data sets. In this study two experiments were carried out: In the first experiment all the data sets were divided into three sets: the ETM+ image and the NDVI that take the scale of near 30m (expect the ETM+ 8 is 14.25m); the DEM, aspect and slope data taking a scale of 25m; the GIS distance data take a scale of 10m. Three times of DT learning and classification using different data sets (RS, RS+DEM, RS+DEM+GIS) were repeated to check whether this set would cause additional misclassification or improve the result.

In the other experiment we select different scales to compare the affection to the DT learning accuracy among data resolutions. According to the data sets, 5 levels of scale (10, 20, 30, 40 and 50) were chosen during each DT and knowledge classification. All the data set (RS image, DEM, and GIS data) were changed into a scale in each classification before the DT learning. The result would display the sensitivity of the DT learning process to the data scales and would specify the best data resolution.

3. RESULTS

In this study we firstly set the DT learning and knowledge engineer for the classification. Secondly the two scale experiments were executed. Then we compared the result of the classifications and selected a best data scale. Finally we achieved the classification using the right data scale.

3.1 Classifications using different data sets

The DT learning and knowledge classification based on a set of spatial data, in which the RS image is the most important. So in the study the RS image was taken as the main data set while other data such as DEM, GIS data were used as assistant information. Results of the three classifications showed the accuracy had an interesting change: the accuracy would increase when we use the DEM as the assistant data, while that would a little decrease when we add the GIS data into the data set. The classification took the highest accuracy of 78.3% using the RS+DEM data and the lowest of 74.7% using RS image only (Figure 4).

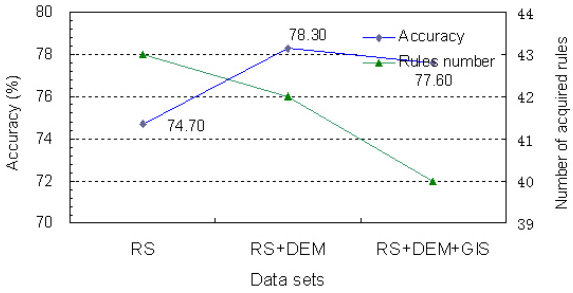

Figure 4. The changing accuracies for the DT learning process and knowledge classification using different data sets.

3.2 DT learning from data sets with different scales

The second experiment showed the affection of data sets with different scales to the DT learning accuracy. Results (Figure 5) showed that the accuracy got the highest value of 80.1% when the data scale is 20m. While the scale increased or decreased, both of the classification accuracy decreased and it would decrease faster as the scale became smaller. However, the number of rules taken by the DT learning kept decreasing as the scale became smaller.

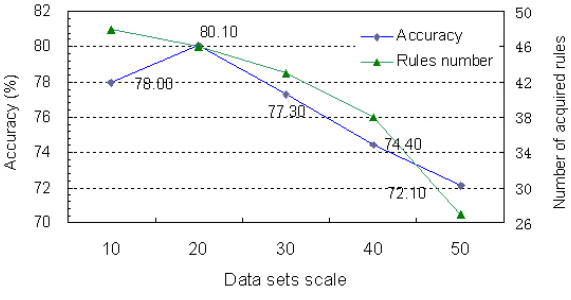

Figure 5. The classification accuracy and the number of rules vary with the different data scales

3.3 Landscape Classification based on the DT Learning

According to the researched affections of the data sets and scales, we chose the data set containing RS image and DEM data for our classification. Meanwhile we rescaled all the data

sets to 20m. Finally the landscape map of the study area was achieved using the DT learning and knowledge classification with the number of rules of 42 and overall accuracy of 80.7% (Figure 6).

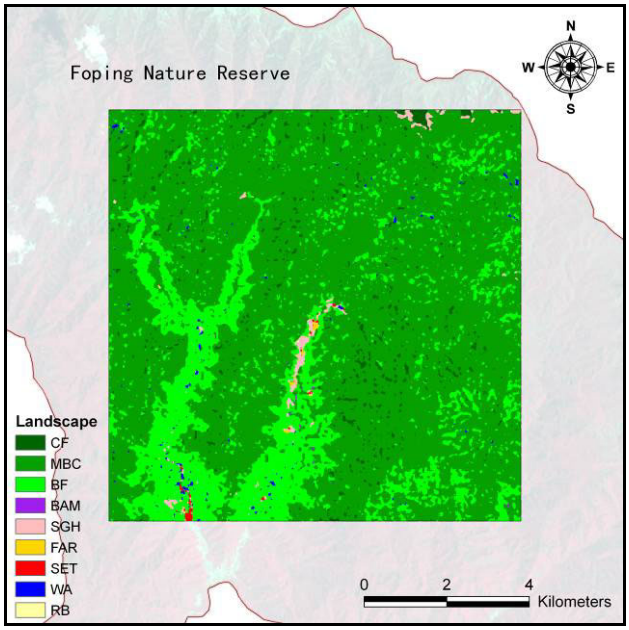

Figure 6. The landscape classification of the study area in Foping NR, China. The used data set took a scale of 20 m. Nine landscape types are conifer forest (CF), mixed broadleaf and conifer forest (MBC), broadleaf forest (BF), bamboo (BAM), shrub/grass/herb (SGH), farmlands (FAR) settlements (SET), water (WA), rock and bare land (RB).

4. DISCUSSION

Former research on the DT learning classification showed that the DT process could handle data sources with different scales (Mahesh and Paul, 2003). However, in this study we found the classification result would be affected by the scale difference. The scale of both partial data and the whole data sets would influence the rules accuracy.

4.1 Data Sets Affection

The data sets in this study mainly contain three different scales. Each data set took certain contribution to the classification. Meanwhile each of these data sets has some errors. When we only used the RS image, we got the classification accuracy of 74.7%. When we used the RS image and the additional DEM data, the classification accuracy could be increased to 78.3%. That showed the DEM data to some extend enhanced the classification. However, when we add the GIS data into the classification resource, the accuracy showed a little decrease. Such a change displays that not all the data with high resolution could offer useful data for DT learning. Maybe the error contained in the data would influence more than the information it provided.

As for this study, the RS image with the DEM data could offer enough information for the DT learning. So at this situation it's not necessary to add the GIS data into the data source. Otherwise it could not only cause additional error to decrease the classification accuracy but also cost computing time and man power.

4.2 Scaling Affection

The scale of the whole data sets could also influence the DT learning and the RS classification. The interesting thing is the classification accuracy didn't increase as the data scale became higher. In our experiment to the Foping NR, the accuracy reached highest when we use the data set with the 20m scale. If the scale became higher, the classification accuracy decreased slowly.

As the data set scale became higher, the number of the rules acquired from the DT learning increased. However, unfortunately not all the rules acquired could be effectively used in the classification. In this study we aimed at classifying the landscape which always displays its characters at the scale more than 10m. To get a better classification using the DT learning, we could change the scale of all the data sets into 20m.

4.3 Effective Classification and Mapping

The DT learning is an effective model used for classification. According to Mahesh (2003), the DT could carry out a better classification than the maximum likelihood classifier when the feature bands number is less than 20. Besides it will take less time to get enough knowledge than neural network classifier. In this study, the DT learning model used for knowledge acquirement got a rule set according to the RS and DEM data and finally we got the landscape map with an acceptable accuracy.

5. CONCLUSION

The expert classifier could use the date set of multi-source besides the RS image and could get acceptable result based on enough knowledge. To solve the problem of acquiring expert knowledge, more and more data mining tools were used in the expert system region. Among these tools, the decision tree learning shows great advantage on knowledge acquirement from multi scale spatial data. More over such knowledge could be convert into rules directly and used for classification.

However, the data scale would influence the accuracy of the knowledge acquired by the DT learning. Each data source could cause certain error besides providing useful information. In this study, we aimed at classifying the landscape distribution. According to the two experiments, RS and DEM data are enough for the DT learning and the best scale was 20m. This result showed it is necessary to select the useful data and the scale for the classification.

REFERENCES

Di K. 2001. Spatial Data Mining and Knowledge Discovery. Wuhan University Press, Wuhan, China, pp. 143-156.

Eric C. B. C., Michael H. S., Craig T., Kathy C., Timothy G. S., and James R. I. 2003. National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier. Remote Sensing of Environment, 85, pp. 316-327.

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

Geoffrey I. W. 1996. Integrating machine learning with knowledge acquisition through direct interaction with domain experts. Knowledge-Based Systems, 9, pp. 253-266.

Huang X. and John R. J. 1997. A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. Engineering & Remote Sensing, 63(10), pp. 1185-1194.

Joseph G., and Gary R. 2002. Expert Systems Principles and Programming, Third Edition. China Machine Press, China, pp. 58-73.

Liu Y., Niu Z., and Wang C. 2005. Research and application of the decision tree classification using MODIS data. Journal of Remote Sensing, 9(4), pp. 405-412.

Mahesh P. and Paul M. M. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86, pp. 554-565.

Mehmed K. 2002. Data Mining Concepts, Models, Methods, and Algorithms. IEEE Press, US, pp. 120-143.

Quinlan J. R. 1993. C4.5: Programs for machine learning. San Mateo: Morgan Kaufmann.

Shi Z. 2002. Knowledge Discovery. Tsinghua University Press, Beijing, China, pp. 21-56.

ACKNOWLEDGEMENTS

Authors of this paper would like to express our sincere thanks to Foping Nature Reserve for their support, which made it possible to successfully implement our field survey during summer 2004. We would also like to thank Jin Xuelin and Dang Gaodi who provided us valuable helps for our field survey.