A SEMI-AUTOMATIC RULE SET BUILDING METHOD FOR URBAN LAND COVER CLASSIFICATION BASED ON MACHINE LEARNING AND HUMAN KNOWLEDGE

H. Y. Gu [a, *], H.T. Li [a], Z.Y. Liu [a], C. Y. Shao [a]

[a] Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, 28 Lianhuachi Road, Beijing, P.R. China - (guhy,lhtao,zjliu)@casm.ac.cn

Commission VI, WG VI/4

KEY WORDS: Land cover classification, Rule set, Machine learning, Human knowledge

ABSTRACT:

Classification rule set is important for Land Cover classification, which refers to features and decision rules. The selection of features and decision are based on an iterative trial-and-error approach that is often utilized in GEOBIA, however, it is time-consuming and has a poor versatility. This study has put forward a rule set building method for Land cover classification based on human knowledge and machine learning. The use of machine learning is to build rule sets effectively which will overcome the iterative trial-and-error approach. The use of human knowledge is to solve the shortcomings of existing machine learning method on insufficient usage of prior knowledge, and improve the versatility of rule sets. A two-step workflow has been introduced, firstly, an initial rule is built based on Random Forest and CART decision tree. Secondly, the initial rule is analyzed and validated based on human knowledge, where we use statistical confidence interval to determine its threshold. The test site is located in Potsdam City. We utilised the TOP, DSM and ground truth data. The results show that the method could determine rule set for Land Cover classification semi-automatically, and there are static features for different land cover classes.

1. INTRODUCTION

Classification rule set is an important method for remote sensing image classification (Forestier, 2012). Rau presented a semiautomatic landslide recognition method using rule set, and validated that the rule set was suitable for various landslide (Rau, 2014). Ziaei presented a rule-based parameter aided with object-based classification approach for extraction of building and roads from WorldView-2 images (Ziaei, 2014).Yu explored the potential role of feature selection in global land-cover mapping (Yu, 2016). Chen measured the effectiveness of various features for thematic information extraction from very high resolution remote sensing imagery (Chen, 2015). However, these methods usually use semiautomatic detection, empirical description and fuzzy function classification. The whole process not only needs supervision, but also requires manual production. Against this background, the next section discusses a semi-automatic rule set building method based on machine learning and human knowledge. The use of machine learning is to build rule sets effectively which will overcome the iterative trial-and-error approach. The use of human knowledge is to solve the shortcomings of existing machine learning method on insufficient usage of prior knowledge, and improve the versatility of rule sets. Urban Land-cover classification test is carried out in order to validate the performance of the method.

2. METHOD

2.1 Rule Set based on Machine Learning

2.1.1 Feature Selection based on Random Forest: The Random Forest (RF) machine learning method is an ensemble classifier developed by Leo Breiman in 2001, based on multiple decision trees. It is a relatively new, non-parametric, data-driven classification method that can create a classification model automatically by learning and training using samples provided by the RS expert, without requiring any prior input (Breiman, 2001). It has the ability to analyze complex features and is robust for noisy and missing data; it is also able to estimate the importance of features and has a faster learning speed and greater accuracy than other similar algorithms that are currently popular (Breiman, 2001).

The RF classifier offers an internal feature evaluation step, through which it is able to estimate the importance of a particular feature, and to subsequently guide the construction of classification rules using significant features only, whereas a general classification method does not offer any form of feature evaluation. It is also able to use a smaller number of features and thus reduce computing time and memory requirements, with no detrimental effect on performance.

The importance of the features is estimated by the RF algorithm, the difference between the current OOB (Out Of Bag) error and the previous OOB error is taken to represent the importance of the variable (Verikas, 2011). Variables with higher values are considered to be more important to the classification than those with lower values. Given a sample subset (s=1,2,…,S) the

* Corresponding author

computation of the importance value (D_j) of feature x_j is as follows:

(a) When s=1, the OOB data L_s^{oob} are classified by decision tree T_s, and the classification number is recorded as N_s^{oob}.

(b) For variable x_j, j=1,2,…,N. When x_j is changed, then L_s^{oob} is also changed and recorded as L_{sj}^{oob}; L_{sj}^{oob} is classified by decision tree T_s and the classification number is recorded as N_{sj}^{oob}.

(c) For s=2,…,S, repeat steps (a) and (b).

(d) The formula for the importance value (D_j) of feature x_j is then: $D_j = 1/S\sum(N_s^{oob} - N_{sj}^{oob})$.

2.1.2 Rule Set Building based on CART: CART is forecasted and classified by constructing binary tree. It has the characteristics of simple model construction, accurate prediction and reusable decision tree rules. This study uses CART for object-based image classification, on the one hand, to test the representative of the features, on the other hand, to reuse the decision tree for other similar image. It includes training and testing steps: (a) Training step. The Gini coefficient of each attribute of the training sample is calculated according to the principle of Gini coefficient gain as the classification condition. After a node generates left and right nodes, the decision tree is generated recursively to divide the left and right nodes, and the decision tree is simplified by pruning method, then the decision tree model is got. (b) Testing step. All the objects are classified by CART decision tree model, and the classification results are obtained.

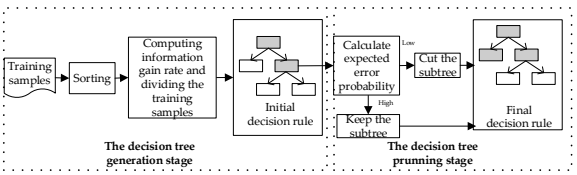


Figure 1. Decision rule based on CART.

2.2 Rule Set based on Human Knowledge

The description and decision rules of eight land-covers are shown in Table 1.

Table 1. The description and rule set of eight land-covers.

	Description	Rule Set
Field	Field is often cultivated for planting crops, which includes cooked field, new developed field and grass crop rotation land. It is mainly for planting crops, and there are scattered fruit trees, mulberry trees or others.	Regular ∩ Planar ∩ Smooth ∩ Dark ∩ Low ∩ adjacentToRoad.
Orchard	Orchard is artificially cultivated for perennial woody and herbaceous crops. It is mainly used for collecting fruits, leaves, roots, stems, etc. It also includes various trees, bushes, tropical crops and fruit nursery, etc.	Regular ∩ Planar ∩ Smooth ∩ Dark ∩ Medium ∩ adjacentToField.
Woodland	Woodland is covered of natural forest, secondary forest and plantation, which includes trees, bushes, bamboo, etc.	Irregular ∩ Planar ∩ Rough ∩ Dark ∩ High ∩ adjacentToField.
Grassland	Grassland is covered of herbaceous plants, which includes shrub grassland, pastures, sparse grassland, etc.	Irregular ∩ Planar ∩ Smooth ∩ Dark ∩ Low∩adjacentToBuilding.
Building	Building includes contiguous building areas and individual buildings in urban and rural areas.	Regular ∩ Planar ∩ Rough ∩ Light ∩ High ∩ adjacentToRoad.
Road	Road is covered by rail and trackless road surface, including railways, highways, urban roads and rural roads.	Regular ∩ Strip ∩ Smooth ∩ Light ∩ Low ∩ adjacentToBuilding.
Bare land	Bare land is a variety of natural exposed surface (forest coverage is less than 10%).	Irregular ∩ Planar ∩ Rough ∩ Light ∩ Low.
Water	Water includes all types of surface water.	Irregular ∩ Planar ∩ Smooth ∩ Dark ∩ Low.

For example, mark rules are shown as follows:

- RectFit (?x, ?y), greaterThanOrEqual(?y, 0.5) -> Regular (?x);
- RectFit (?x, ?y), lessThan(?y, 0.5) -> Irregular (?x);
- LengthWidthRatio(?x, ?y), greaterThanOrEqual(?y, 1) -> Strip(?x);
- LengthWidthRatio(?x, ?y), lessThan (?y, 1) -> Planar(?x);

This means RectFit of an object >0.5 denotes Regular shape, where <0.5 denotes Irregular shape. The thresholds are obtained by a statistical confidence interval approach.

2.3 Statistical Confidence Interval

In statistics, a confidence interval is a type of interval estimate of population parameter constructed by the sample statistic. Two-side confidence limits from a confidence interval and one-side limits are referred to as lower/upper confidence bounds (or limits). The affect factors include the size of samples and the confidence level. In the case of a fixed confidence level, the more the samples, the narrower the confidence interval. In the case of a fixed samples, the higher the confidence level, the wider the confidence interval. The confidence interval is defined as:

[M-N*STD, M+N*STD]

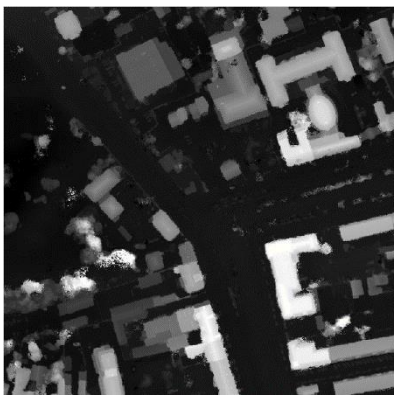
Where, M is the mean of the sample, STD is the standard deviation of the sample, N is used as the critical value.

3. EXPERIMENT AND ANALYSIS

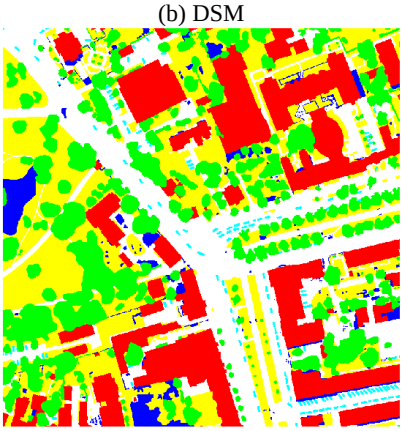
3.1 Data

The data set is in the city of Potsdam. We used true orthophoto (TOP) data with four channels red, green, blue, infrared, and the DSM and ground truth (ISPRS, 2017). The TOP and DSM are used for classification, the ground truth is used for sample selection.

(a) True orthophoto (TOP) data



(b) DSM



(c) Ground truth data

Figure 2. Data set in Potsdam

3.2 Experiment

(1) Segmentation. Firstly, we use ArcGIS to make the ground truth data as a vector constraints. Then we use eCognition for multi-resolution segmentation. The trial-and-error method is adopted to find an approximate and reasonable scale parameter, where the scale is set to 100, the shape factor weight is 0.2 and compactness factor weight is 0.8.

(2) Feature Selection. Sixteen features (e.g., ratio, mean, Normalized Difference Water Index, Normalized Difference Vegetation Index, homogeneity, and brightness) are selected, and then are sorted using RF. The feature importance of Potsdam is shown in figure 3.

Variable	Score
MEAN_DSM	100.00
NDVI	99.31
NDWI	98.29
RATIO_GREEN	78.37
BRIGHTNESS	55.14
MAX__DIFF_	46.99
MEAN_NIR	45.99
STANDARD_DEVIATION_BLUE	30.42
AREA	20.45
LENGTH_WIDTH	15.17
DENSITY	12.00
SHAPE_INDEX	11.62
LENGTH	9.61
GLCM_MEAN__ALL_DIR__	5.59
GLCM_HOMOGENEITY__ALL_DIR__	4.78
RECTANGULAR_FIT	4.25

Figure 3. The feature importance of Potsdam

(3) Decision Rules Building. The initial decision rules are built using CART. Which is shown in figure 4 and table 2.

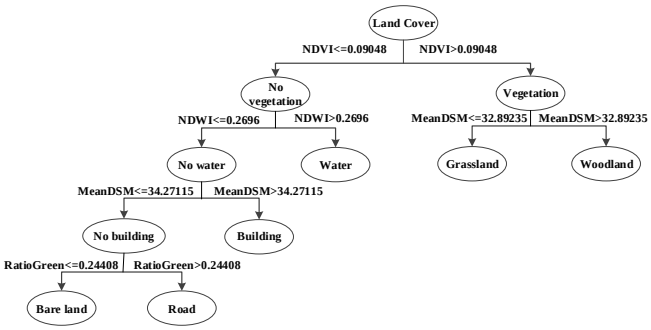


Figure 4. The decision rules of Potsdam

Table 2. Decision rules

Class	Decision rules
Woodland	NDVI>0.09048&MeanDSM>32.89235
Grassland	NDVI>0.09048&MeanDSM<=32.89235
Building	NDVI<=0.09048 &NDWI <=0.269& MeanDSM>34.27115
Road	NDVI<=0.09048 &NDWI <=0.269& MeanDSM<=34.27115& RatioGreen>0.24408
Water	NDVI<=0.09048 & NDWI>0.269
Bare land	NDVI<=0.09048 &NDWI <=0.269& MeanDSM<=34.27115& RatioGreen<=0.24408

(4) Decision Rules validation. The initial rule is validated based on human knowledge and membership function, where we use statistical confidence interval to determine its threshold. For example, the confidence interval of building is shown in table 2.

Table 3. The confidence interval of building

Object types	Features	Member ship function	Confidence interval
building	DSM	>	$[DSM_{M-N*std}, DSM_M]$
	ratioG	<	$[ratioG_M, ratioG_{M+N*std}]$

(5) Classification. The image are classified using CART, and then are validated using human knowledge and membership function. The

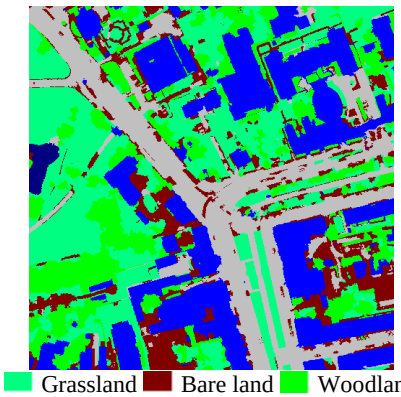


Figure 5. Classification results of Potsdam

3.3 Analysis

An accuracy assessment was carried out. A sample-based error matrix is created and used for performing accuracy assessment. In GEOBIA, a sample refers to an object. The error matrix for the test area is shown in Figure 16. The user's accuracy, producer's accuracy, overall accuracy and Kappa coefficient is shown in Table 4.

Table 4. Overall accuracy.

	grassland	road	woodland	building	water	bareland	overall	UA
grassland	30	0	1	0	0	0	31	96.77
road	0	28	0	0	1	0	29	96.55
woodland	0	0	29	0	0	0	29	100
building	0	0	0	30	0	0	30	100
water	0	0	0	0	5	0	5	100
bareland	0	2	0	0	0	8	10	80
overall	30	30	30	30	6	8	134	
PA%	100	93.33	96.67	100	83.33	100		

OA=97.01%，Kappa=0.96

The overall accuracy is 97.01%, and the kappa coefficient is 0.96. Our method yields improvements as it depends on decision rule based on machine learning and human knowledge. This is based on the initial decision rules and the validation process, and some obvious classification errors may be corrected already within the following validation step.

4. CONCLUSION

This study has put forward a rule set building method for Land cover classification based on human knowledge and machine learning. The use of machine learning is to build rule sets effectively which will overcome the iterative trial-and-error approach. The use of human knowledge is to solve the shortcomings of existing machine learning method on insufficient usage of prior knowledge, and improve the versatility of rule sets. A two-step workflow has been introduced, firstly, an initial rule is built based on Random Forest and CART decision tree. Secondly, the initial rule is analyzed and validated based on human knowledge, where we use statistical confidence interval to determine its threshold. The test site is located in Potsdam City. We utilised the TOP, DSM and ground truth data. The results show that the method could determine rule set for Land Cover classification semi-automatically, and there are static features for different land cover classes.

Nevertheless, the method is still in the process of development and improvement. Further in-depth studies may be required to (a) improve and refine rule set using human knowledge, (b) investigate the factors influencing classification, such as the spatial scale, the segmentation method employed, and the choice of samples, and (c) to investigate the automation of the method.

ACKNOWLEDGEMENTS

This research was funded by: (1) the National Natural Science Foundation of China (Project Nos. 41371406, 41471299, and 41671440); (2) Central Public-interest Scientific Institution Basal Research Fund (Project Nos. 7771508, 777161101, and 777161102).

REFERENCES

Breiman, L, Cutler, A. Random Forests; Available online: http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (accessed on 5 March 2017).

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5-32.

Chen X, Fang T, Huo H, Li D R., 2015. Measuring the Effectiveness of Various Features for Thematic Information Extraction From Very High Resolution Remote Sensing Imagery. IEEE Transactions On Geoscience And Remote Sensing, 53(9):4837-4851.

Forestier G, Puissant A, Wemmert C, et al., 2012. Knowledge-based region labeling for remote sensing image interpretation. Computers, Environment and Urban Systems, 36(5): 470–480.

ISPRS Test Project on Urban Classification, 3D Building Reconstruction and Semantic Labeling. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html. (accessed on 5 March 2017).

Rau J Y, Jhan J P, Rau R J., 2014. Semiautomatic Object-Oriented Landslide Recognition Scheme From Multisensor Optical Imagery and DEM. Geoscience & Remote Sensing IEEE Transactions on, 52(2):1336-1349.

Verikas, A., Gelzinis, A., Bacauskiene, M. Mining data with random forests: A survey and results of new tests. Pattern Recognit. 2011, 44, 330-349.

Yu L, Fu H H, Wu B, Clinton N, Gong P., 2016. Exploring the potential role of feature selection in global land-cover mapping. International journal of remote sensing, 37(23): 5491-5504.

Ziaei Z, Pradhan B, Mansor S B., 2014. A rule-based parameter aided with object-based classification approach for extraction of building and roads from WorldView-2 images. Geocarto International, 29(5):554-569.