Improved Building Roof Type Classification Using Correlation-Based Feature Selection and Gain Ratio Algorithms

M. Norman, H. Z. M. Shafri, Biswajeet Pradhan and B. Yusuf

Abstract Of late, application of data mining for pattern recognition and feature classification is fast becoming an essential technique in remote sensing research. Accurate feature selection is a necessary step to improve the accuracy of classification. This process depends on the number of feature attributes available for interactive synthesis of common characteristics that discriminate different features. Geographic object-based image analysis (GEOBIA) has made it possible to derive varieties of object attribute for this purpose; however, the analysis is more computationally intensive. The aim of this study is to develop feature selection technique that will provide the most suitable attributes to identify different roofing materials and their conditions. First, the feature importance was evaluated using gain ratio algorithm, and the result was ranked, leading to selection of the optimal feature subset. Then, the quality of the selected features was assessed using correlation-based feature selection (CFS). The classification results using SVM classifier produced an overall accuracy of 83.16%. The study has shown that the ability to exploit rich image feature attribute through optimization process improves accurate extraction of roof material with greater reliability.

M. Norman · H. Z. M. Shafri (✉) · B. Pradhan · B. Yusuf
Department of Civil Engineering, Faculty of Engineering,
Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia
e-mail: hzms04@gmail.com

B. Pradhan
e-mail: biswajeet24@gmail.com

H. Z. M. Shafri · B. Pradhan
Geospatial Information Science Research Centre (GISRC), Faculty of Engineering,
Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia

M. Norman
Department of Surveying Science and Geomatics, Faculty of Architecture,
Planning and Surveying, Universiti Teknologi MARA (Perlis),
02600 Arau, Perlis, Malaysia
e-mail: ayunorman@gmail.com

B. Pradhan (ed.), *GCEC 2017*, Lecture Notes in Civil Engineering 9,
https://doi.org/10.1007/978-981-10-8016-6_62

Keywords Optimization building classification · Feature selection
Gain ratio · Correlation-based selection (CFS)

1 Introduction

In remote sensing applications, roof identification and extraction is challenging due to similarities in geometric pattern, variation in physical and chemical properties and spatial heterogeneity. Geographic object-based image analysis (GEOBIA) has been widely used as it can generate numerous feature attributes of the image object for better feature classification. OBIA exploits the spectral, shape and texture characteristics of the images to partition it into features in vector form. It has been reported to provide better feature extraction capabilities compared to traditional pixel-based image classification techniques that accord it the widespread popularity among the remote sensing practitioners [1, 2]. Moreover, its feature attribute extraction oriented output facilitates data mining applications for pattern recognition [3]. To achieve a high degree of automation process, combination of quantitative attribute-based feature selections and machine learning classification are very significant.

Due to the high frequency of image objects and effect if illuminations versus sensor look angle, particularly in urban area, it is always difficult to group pixels belonging to the same object within a single image segment. This problem is more pronounced when unblended segmentation parameters are used [4, 5]. This causes numerous numbers of segments especially in high-resolution imagery making analysis of feature more computationally intensive and prone to error in classification.

Several feature selection algorithms have been used in the past to select relevant feature for different applications. Novack et al. [6] applied four different types of feature selection algorithms: Info-Gain, Relief-F, Fast Correlation-Based Filter (FCBF) and Random Forest, to provide ranking of variable importance. They found out that the features selected from the four algorithms are the same for each class group. This is because most of the algorithms consider not only the relevance of the features but also redundancy among each other. Recently, Li et al. [7] applied correlation-based feature selection (CFS) to measure the quality of a subset of features for land-use land-cover classification. The result shows that combining feature selection with classifiers such as Random Forest, the technique improved the accuracy of the classification. In this research, we present an effective means of extracting roof materials and their conditions using multispectral World-View 3 imagery through the process of feature attribute exploration.

2 Study Area and Datasets

This study was carried out over Universiti Putra Malaysia (UPM) and its surrounding, covering about 2.7 km^2 area (Fig. 1). This comprises residential type building setting with different roof materials including metal roof, concrete roof and asbestos roof, which are categorized according to their materials and conditions (new and old). The landscape presents mixture pervious and impervious surfaces such as water, road, trees, grass and bare land. High spatial resolution World-View 3 with panchromatic image with 0.31 spatial resolution and multispectral image with 1.24 m resolution were used for the investigation. This satellite was successfully launched on 13 August 2014. It is the first multi-payload, super-spectral, high-resolution commercial satellite sensor operating with eight multispectral bands (coastal, blue, green, yellow, red, red-edge, NIR 1 and NIR 2).

3 Methodology

3.1 Preprocessing and Segmentation

The World-View 3 image was corrected for radiometric and atmospheric effects to create a more faithful representation of the original scene. Subsequently, the image was segmented using the multiresolution segmentation algorithm in eCognition v9.1 [8]. With the define segmentation parameters, over-segmentation occurred leaving some features being represented by many image objects. This was corrected

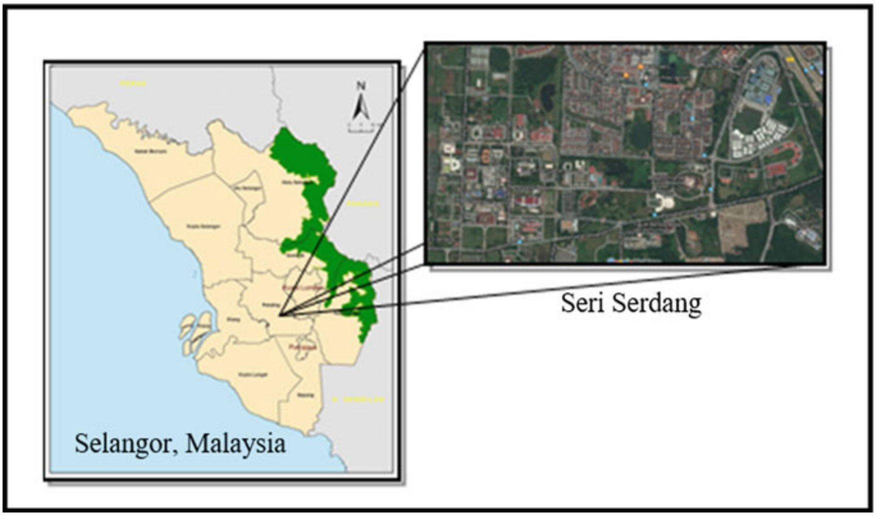


Fig. 1 Study area

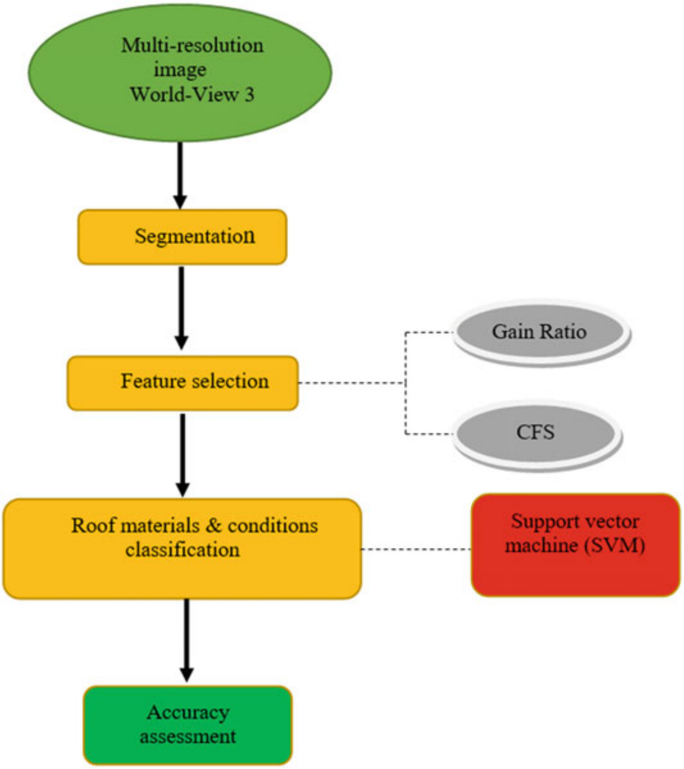

Multi-resolution image World-View 3
Segmentation
Feature selection
Gain Ratio
CFS
Roof materials & conditions classification
Support vector machine (SVM)
Accuracy assessment

Fig. 2 Framework of the overall process

by merging contiguous segments into larger segment representing object entity based on the spectral similarity threshold. Schematic flow of data processing is presented in Fig. 2.

Five levels of segmentation parameters were defined using five different scale parameters (60, 70, 80, 90 and 100). Similarly, the shape and compactness parameters were set as 0.1, 0.3, 0.5, 0.7 and 0.9. These parameters were optimized using Taguchi method. The results of the experiment obtained the optimal parameter combinations 80:0.7:0.1 for scale, shape and compactness, respectively.

3.2 Features Selection

The advantage of OBIA is the ability to extract multiple attributes about the image objects, which allows intricate data exploration for object classification. However, processing the huge attribute data generated is computationally intensive. So, gain ratio (GR) algorithm (Eq. 1) [9] was used to assess the importance of the features attributes so as to select most important one in order minimizes the computational requirements. GR eliminates bias towards tests with many outcomes (large attribute values) by measuring information. The information gain measures interaction of attribute selection by employing decision tree, which computes the difference among the predicted information requirement, and then classifies a tuple in tuples for update information prerequisite. The greater the gain ratio, the more significant it represents features [10]. GR is expressed as

$$\text{Gain Ratio}\,(\text{A}) = \frac{\text{Gain}\,(\text{A})}{\text{Split Info}_{\text{A}}(\text{D})} \tag{1}$$

Having obtained the degree of importance of feature attributes with GR, correlation-based feature selection (CFS) (Eq. 2) was utilized to evaluate the quality of the features subset from which only the significant features for roof materials and conditions discrimination were selected. CFS simply evaluates the worth of a set of features using a heuristic assessment function based on the correlation of the features. A good feature subset contains features highly correlated with classes and highly uncorrelated to each other [11].

$$\text{Merits} = \frac{k\bar{r}_{\text{cf}}}{\sqrt{k + k(k-1)}r_{\overline{\text{ff}}}} \tag{2}$$

where f indicates the feature, c is the class, $\bar{r}_{\text{cf}}$ denotes the mean feature correlation with classes, $r_{\overline{\text{ff}}}$ indicates the average feature intercorrelation and k denotes the number of the attributes in the subset.

3.3 Classification and Accuracy Assessment

The image object represents unclassified features. In this study, support vector machine (SVM) was used to assign feature class to the image objects. SVM is based on statistical computational process that identifies the optimal hyperplane as a decision function in high-dimensional space [12]. The previous study shows that the number of features selected will determine the accuracy of SVM classification [13]. 12 classes, namely, tree, grass, water, road, shadow, bare land, new metal, old metal, new concrete, old concrete, new asbestos and old asbestos were extracted in the process. The accuracy of the classification was evaluated using standard confusion matrix [14].

4 Result of Roof Materials and Conditions Discrimination

The results (Fig. 3a–c) indicate the rank of roof materials and conditions features according to gain ratio value from three categories, spectral, shape and texture. They have been ranked from the highest value to the lowest.

Table 1 shows that 11 features selected as the most suitable for classification.

The accuracy was calculated using standard confusion matrix method, whereby the results for classification were represented by this matrix (Fig. 4). Thus, the overall accuracy can be computed as well as individual class accuracy.

Figure 5a shows that new metal misclassified with old asbestos and after suitable features applied, the misclassification improves to be as new metal for the whole object (Fig. 5b). Meanwhile, there are three objects misclassified in Fig. 5c. Feature selection approach successfully solves the misclassification problem, whereby each object has been classified into its true class, such as new concrete re-classified as old concrete, old metal as new concrete and old metal as new metal (Fig. 6).

Result shows the classification of roof materials and conditions within UPM Serdang main campus and its surrounding area using SVM classifier. It proved that the result is much better after using the selected features to classify the image. As a result, overall accuracy for roof materials and conditions classification using SVM classifier is 83.16% and kappa coefficient is 0.81. After classification using selected features applied to the images, the results obviously show the improvement of misclassification.

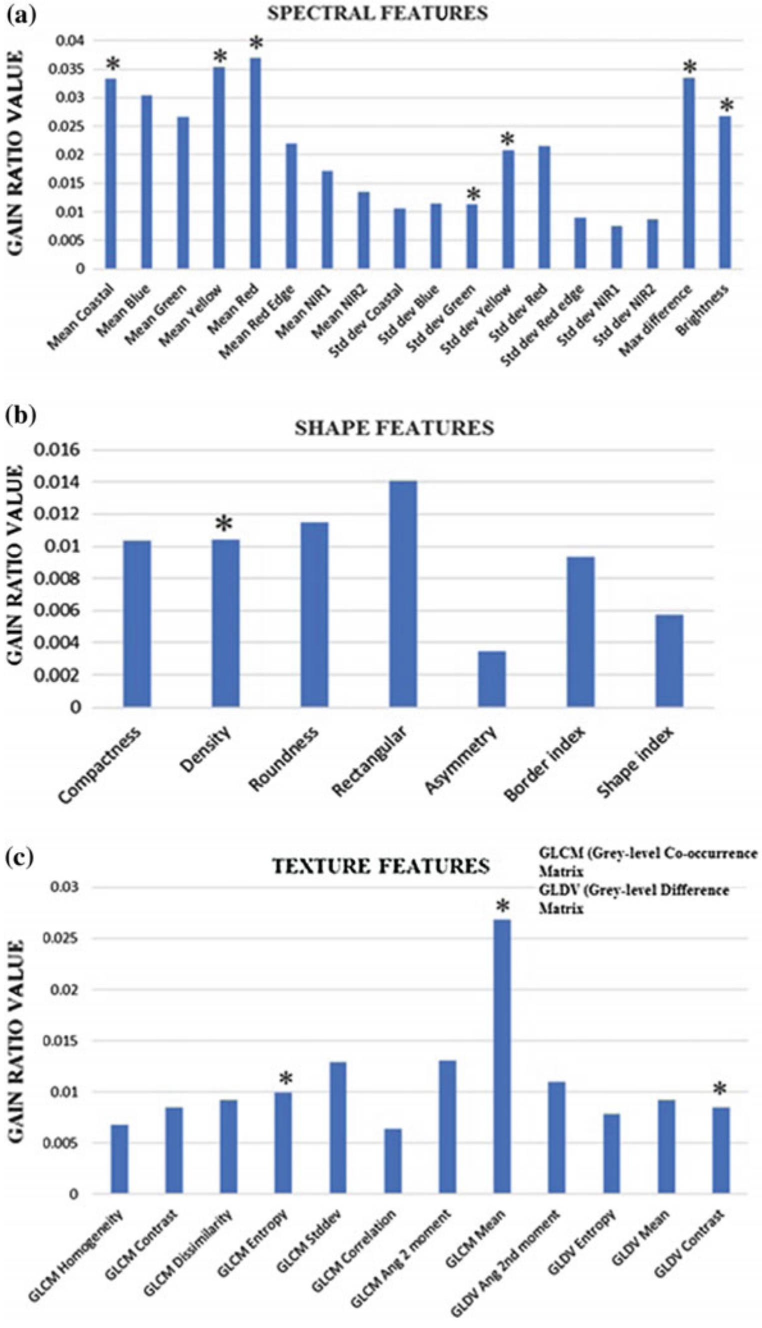
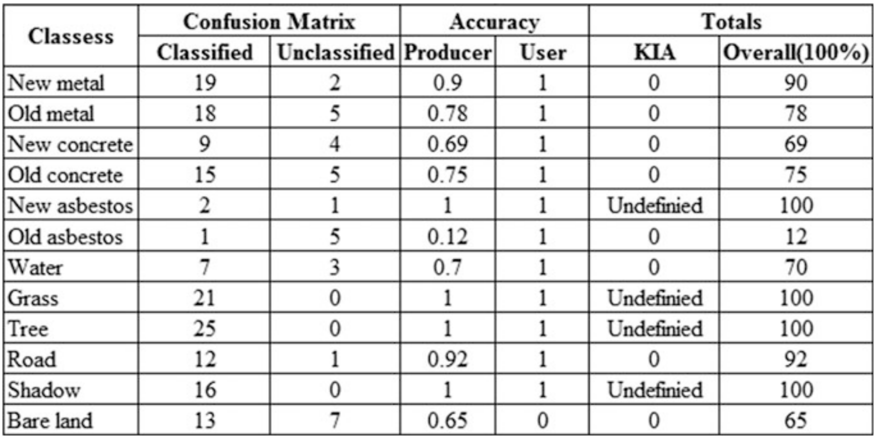


Fig. 3 Relative contributions of roof materials and conditions features based on gain ratio index

Table 1 Selected features

No.	Feature	Category
1	Mean coastal	Spectral
2	Mean yellow	Spectral
3	Mean red	Spectral
4	Brightness	Spectral
5	Maximum difference	Spectral
6	Standard deviation green	Spectral
7	Standard deviation yellow	Spectral
8	Density	Shape
9	GLCM entropy	Texture
10	GLCM mean	Texture
11	GLDV contrast	Texture



Classess	Confusion Matrix		Accuracy		Totals	
	Classified	Unclassified	Producer	User	KIA	Overall(100%)
New metal	19	2	0.9	1	0	90
Old metal	18	5	0.78	1	0	78
New concrete	9	4	0.69	1	0	69
Old concrete	15	5	0.75	1	0	75
New asbestos	2	1	1	1	Undefinied	100
Old asbestos	1	5	0.12	1	0	12
Water	7	3	0.7	1	0	70
Grass	21	0	1	1	Undefinied	100
Tree	25	0	1	1	Undefinied	100
Road	12	1	0.92	1	0	92
Shadow	16	0	1	1	Undefinied	100
Bare land	13	7	0.65	0	0	65

Fig. 4 Accuracy for each class

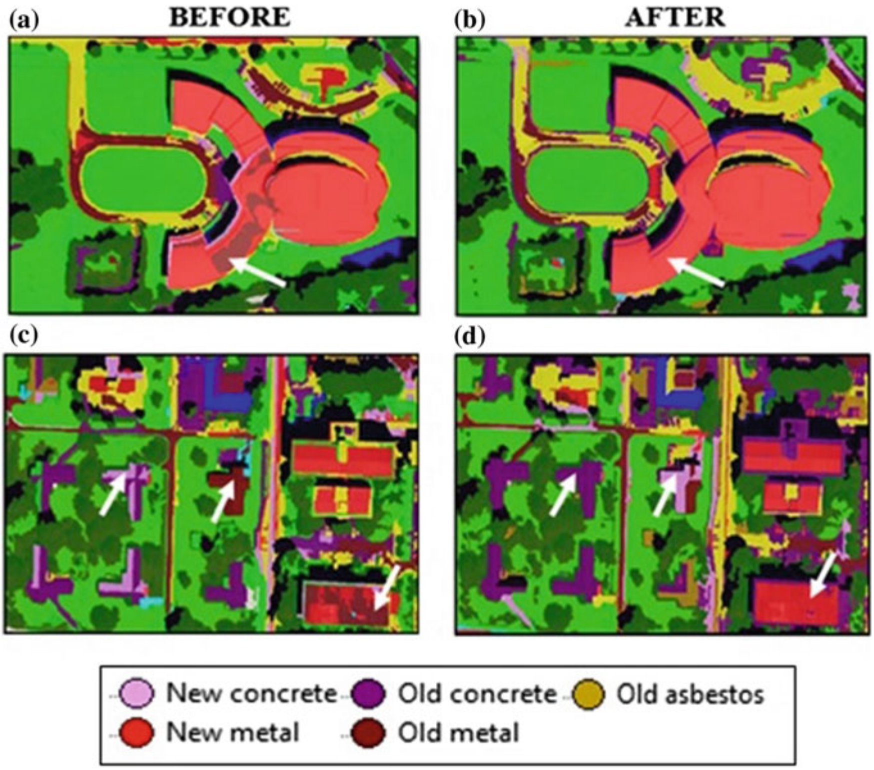


Fig. 5 Example of an improved classification for roof materials and conditions after suitable features applied

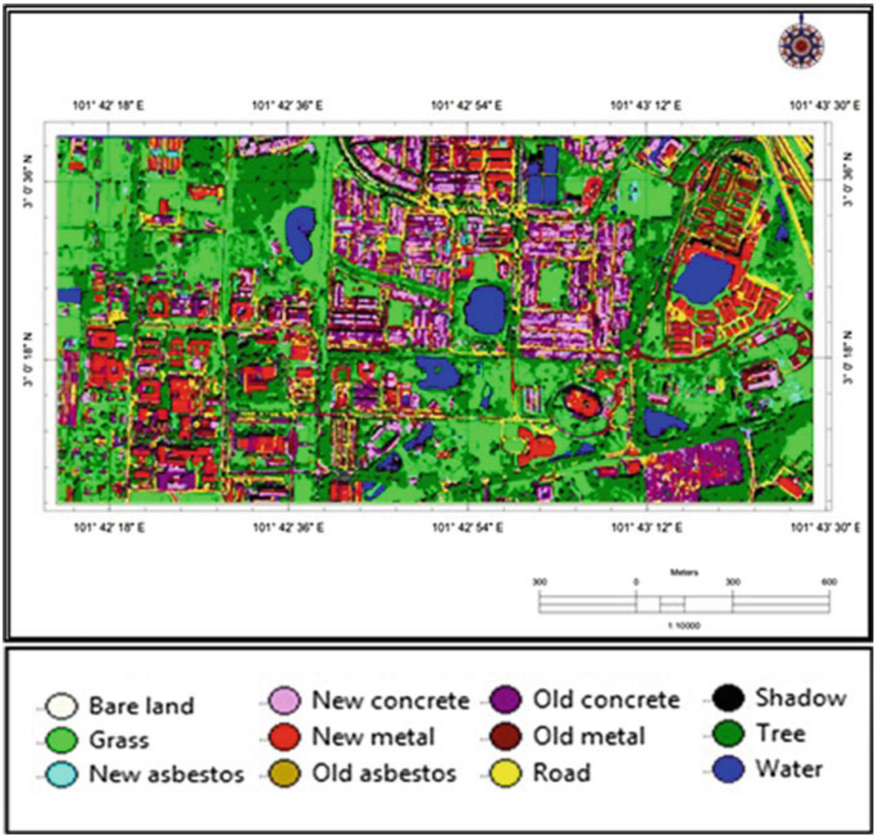


Fig. 6 Distribution of roof materials and conditions using SVM classifier

5 Conclusion

This study aimed to develop feature selection technique to verify the most suitable features to be used for the discrimination of roofing materials and conditions based on different feature algorithms available. The performance assessment was done, giving satisfied accuracy and error confusion matrix. Experimental results demonstrate that SVM based on significant features can improve the quality of classification and reduce the misclassification especially between road and concrete, shadow and water, old concrete and old metal as well. Those misclassifications are due to the similar spectral reflectances between each material.

As a result, the systematic feature selection approach can significantly contribute to roof materials and condition discrimination besides increase its accuracy. The application of gain ratio algorithm in this study is effective for reducing the misclassification problem of roof material classification.

Future research should be conducted by comparing several features selection algorithms, and further assessment of different classifiers is required.

Acknowledgements The authors would like to acknowledge the research fundings provided by the Ministry of Education Malaysia under the FRGS grant (03-02-14-1529FR, Vot no: 5524613), Universiti Putra Malaysia (UPM) under the Geran Putra Berimpak (GPB) grant (Vot no: 9543100), Universiti Teknologi MARA (Perlis), Malaysia and Ministry of Education (MOE) Malaysia.

References

1. Blaschke, T., et al.: Geographic object-based image analysis—towards a new paradigm. ISPRS J. Photogramm. Remote Sens. **87**, 180–191 (2014)
2. Radoux, J., Bogaert, P.: Accounting for the area of polygon sampling units for the prediction of primary accuracy assessment indices. Remote Sens. Environ. **142**(February), 9–19 (2014)
3. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**(3), 131–156 (1997)
4. Kim, M., Warner, T.A., Madden, M., Atkinson, D.S.: Multi-scale GEOBIA with very high spatial resolution digital aerial imagery: scale, texture and image objects. Int. J. Remote Sens. **32**(10); **1161**(10), 2825–2850 (2011)
5. Zhang, X., Xiao, P., Song, X., She, J.: Boundary-constrained multi-scale segmentation method for remote sensing images. ISPRS J. Photogramm. Remote Sens. **78**(May), 15–25 (2013)
6. Novack, T., Esch, T., Kux, H., Stilla, U.: Machine learning comparison between WorldView-2 and QuickBird-2-simulated imagery regarding object-based urban land cover classification. Remote Sens. **3**(10), 2263–2282 (2011)
7. Li, D.T.M., Ma, L., Blaschke, T., Cheng, L.: A systematic comparison of different object based classification techniques using high spatial resolution imagery in agricultural environments. Int. J. Appl. Earth Obs. Geoinf. **49**, 87–98 (2016)
8. Weise, C.: eCognition Essentials, pp. 1–2 (2016)
9. Quinlan, J.R., Improved use of continuous attributes in C4. 5. l Artif. Intell. Res. Artif. Intell. Res. **4**, 77–90 (1996)
10. Ma, L., et al.: Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. Int. J. Geo-Inform. **6**(51), 1–21 (2017)
11. Hall, M.A., Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans. Knowl. Data Eng. **15**(6), 1437–1447 (2003)
12. Bazi, Y., Melgani, F.: Toward an optimal SVM classification system for hyperspectral remote sensing images. Geosci. Remote Sens. IEEE Trans. **44**(11), 3374–3385 (2006)
13. Hamedianfar, A., Shafri, H.Z.M.: Development of fuzzy rule-based parameters for urban object-oriented classification using very high resolution imagery. Geocarto Int. **29**(3), 268–292 (2014)
14. Congalton, R.G., Green, K.: Assessing the Accuracy of Remotely Sensed Data: Principles and Practices (2009)