DATA MINING FOR KNOWLEDGE DISCOVERY FROM OBJECT-BASED SEGMENTATION OF VHR REMOTELY SENSED IMAGERY

K. Djerriri[a,b,*], M. Malki[b]

[a] Division of Earth Observation, Centre for Spatial Techniques, Arzew, Oran, Algeria, kdjerriri@cts.asal.dz
[b] EEDIS Laboratory, University of Sidi Bel Abbes, Algeria, malki@univ-sba.dz

Commission VI, WG VI/4

KEY WORDS: Object-Based Image Analysis, Data Mining, Classification rule Induction, Genetic Algorithm

ABSTRACT:

The success of the object-based image analysis (OBIA) paradigm can be attributed to the fact that regions obtained by means of segmentation process are depicted with a variety of spectral, shape, texture and context characteristics. These representative objects-attributes can be assigned to different land-cover/land-use types by means of two options. The first is to use supervised classifiers such as K-nearest neighbors (KNN) and Support Vector Machine (SVM), the second is to create classification rules. Supervised classifiers perform very well and have generally higher accuracies. However one of their drawbacks is that they provide no explicit knowledge in understandable and interpretable forms. The building of the rule set is generally based on the domain expert knowledge when dealing with a small number of classes and a small number of attributes, but having a dozens of continuously valued attributes attached to each image object makes it a tedious task and experts quickly get overwhelmed and become totally helpless. This is where data mining techniques for knowledge discovering help to understand the hidden relationships between classes and their attached attributes. The aim of this paper is to highlight the benefits of using knowledge discovery and data-mining tools, especially rule induction algorithms for useful and accurate information extraction from high spatial resolution remotely sensed imagery.

1. INTRODUCTION

In the last decade, Object-based methods have been proved to be very useful to work with high spatial resolution remotely sensed data, by saving GIS technicians from digitizing hundreds of objects by hand. The object-based image analysis can be seen as a step toward decreasing the semantic gap between pixel-based low-level spectral features and high-level semantic concepts in the images, and that by handling image primitives as objects instead of individual pixels. Objects are derived from input image by means of image segmentation, which is the process of partitioning images into segments by grouping neighboring pixels with similar spectral characteristics.

The regions obtained by means of segmentation process can be depicted with a variety of spectral, shape, texture, context characteristics and neighborhood relations to other objects. These representative objects-attributes enable a later classification into object classes.

Rule-based classification technique has been proved to be powerful tool, where it is implemented in most widely used remote sensing image classification software, such as eCognition and ENVI Feature Extraction.

The aim of this paper is to highlight the benefits of using some available knowledge discovery and data-mining techniques, especially rule induction algorithms for useful and accurate information extraction from high spatial resolution remotely sensed imagery. The contribution of this paper is twofold. First, we present the results of many data mining algorithms, adapted to the classification of VHR image of an urban area. The second contribution lies in the using of the discovered classification rules to identify appropriate concepts, attributes and their domain values for the building of Ontology of urban objects, which models domain knowledge in a formal, machine understandable and sharable way.

2. METHODS

2.1 Image segmentation and Feature extraction

The first step in OBIA is to generate segments from input image by means of image segmentation (Blaschke, 2010). Up to now, a vast amount of methods and algorithms were developed especially for remotely sensed imagery or adopted from other fields. The ideal segmentation results should correspond to the real-world objects. However, the problem of splitting up the input image into too few (under-segmentation) or too many regions (over-segmentation) constantly occurs during the process; therefore, appropriate segmentation techniques need to be selected and carefully conducted. Based on the idea that not all real world objects occur at the same or a similar detail level, many studies reported that Multi-scale models are the most widely used for segmenting VHR images (Neubert and Herold, 2008). Those methods make object extraction of various scales possible by generating segmentation results from finer to coarser segmentation by merging adjacent regions to different levels.

The segmentation task in our research was performed in ENVI feature extraction. A Multi-scale edge-based segmentation algorithm introduced by (Xiaoying, 2007.) is implemented in this software tool. This method is very fast and only requires

[*] Corresponding author

one input parameter (scale level). Within the same software it is possible to compute 26 descriptors for each image segment, 14 for shape, 4 spectral for each spectral band, 4 for texture and 4 for color space and band ratio, respectively. These descriptors are Total area, Length of all boundaries, Compactness, Convexity, Solidity, Roundness, Factorform, Main Direction, Major Axis length, Minor Axis length, Number of holes, HOLES RATE, MINBAND_x (minimum value of the pixels comprising the region in band x), MAXBAND_x, AVGBAND_x, STDBAND_x (standard deviation value of the pixels comprising the region in band x), TX_RANGE (average data range of the pixels comprising the region inside the texture kernel), TX_MEAN, TX_VARIANCE, TX_ENTROPY, HUE , SATURATION, Intensity (ITT Visual Information Solutions. 2007).

2.2 Data Mining

Knowledge discovery through data mining have known a great success during the past few years. Here, we are interested to discover classification rule for useful and accurate information extraction from high spatial resolution remotely sensed imagery. Thus a set of algorithms have been implemented and tested on the object obtained earlier.

2.2.1 Genetic Algorithm (GA)

GA is a global optimization technique, which can be used to generate high predictive and comprehensible classification rule. A survey of genetic algorithms for data mining and knowledge discovery can be found in (Freitas, A. A., 2003).

In our case each solution encoded by the GA corresponds to optimized intervals of selected relevant attributes, linked with land-cover types for forming classification rules. These rules take the general following format:

IF Atrribute$_1 \in$ Interval$_1$ AND Atrribute$_2 \in$ Interval$_2$ …IF Atrribute$_j \in$ Interval$_j$ THEN Object is Class$_i$

A single rule can be coded as one chromosome consisting of n genes, where each gene corresponds to a segment of 3 elements encoding an attribute or a condition in the condition part and n the total number of attributes. Each segment consists of a bi-valued flag flagi, where value of 0 corresponds to the presence of the ith attribute in the rule and value of 1 to the absence of the attribute, and two continuous cutoff values, the lower bound Lbi and the Upper bound Ubi defining the rangeof ith attribute for the class being in process.
For example the sequence (0.5,0.8,0 ,10.0,20.0,1 ,5.0,7.0,0 ,2.0, 6.0) encodes the rule IF attr$_1 \in$ [0.5,0.8] AND attr$_3 \in$ [5.0,7.0] THEN Object is Class$_i$

Three (03) main genetic operators are used: selection, crossover, and mutation. The selection is a process in which the fittest rules have higher chance of being carried forward to the next generation. Crossover allows information to be exchanged, where the operator chooses a point to be selected on the parent solution sequences then information is swapped between the two individuals, rendering two child solutions. Mutation is used to randomly choose a member of the population and to change one randomly chosen element in its sequence representation. After the processes of selection, crossover, and mutation have been applied to the initial population, a new population will have formed following the replacement step. After replacement, the new population will be evaluated based on its fitness in the next evolution. This process of selection, crossover, mutation, and replacement is continued until a fixed number of generations have been reached or some form of convergence criterion has been met.

A central instrument in a genetic algorithm is the fitness function. Ideally the discovered rules should satisfy two criteria: predictive accuracy and comprehensibility, so these objectives are combined into a single objective fitness function.

Let a rule be of the form: IF A THEN C, where A is the antecedent (a conjunction of conditions) and C is the consequent (predicted class). The class predicted for an example is C if and only if the example satisfies the rule antecedent. The predictive performance of a rule can be summarized by four (04) cases, sometimes called a confusion matrix (Freitas, A. A., 2003).

- True Positives (TP): Number of examples satisfying A and C
- False Positives (FP): Number of examples satisfying A but not C
- False Negatives (FN): Number of examples not satisfying A but satisfying C
- True Negatives (TN): Number of examples not satisfying A nor C

A very simple way to measure the predictive accuracy of a rule is to compute the confidence factor (CF) of the rule, defined as:

$$CF = TP / (TP + FP) \quad (1)$$

It is possible to measure the predictive accuracy of a rule by taking into account not only its CF but also a measure of how "complete" the rule. The rule completeness (comp) measure, denoted Comp, is computed by the formula

$$Comp = TP / (TP + FN) \quad (2)$$

In order to combine the CF and Comp measures we can define a fitness function such as:

$$Fitness = CF * Comp \quad (3)$$

This fitness function does not evaluate the comprehensibility of the rule. A simple measure of comprehensibility is the Simplicity (simp), which in general, its value is inversely proportional to the number of conditions in the rule antecedent.

If a rule has at most L condition, the comprehensibility of the rule R can be defined as:

$$Simp = L-h / L-1 \quad (4)$$

Where h is the length of the rule R and simp is normalized to take on values in the range 0..1.

Finally, the fitness function used by our system is defined as:

$$\text{Fitness} = \text{CF} * \text{Comp}*\text{Simp} \qquad (5)$$

2.2.2 Genetic Programming (GP) for Symbolic Classification

The goal of GP, as its name implies, is the evolution of computer programs or mathematical expressions instead of sequence of values as for GA (Espejo et al., 2010). GP individuals are usually seen as trees, where leaves correspond to terminal symbols (variables and constants) and internal nodes correspond to non-terminals (operators and functions). The set of all the non-terminal symbols allowed is called the function set, whereas the terminal symbols allowed constitute the terminal set. An example of tree that represents the expression for calculating normalized difference vegetation index is illustrated in the figure 1. In our case the terminal symbols are the attributes of objects and the internals nodes are mathematical operators such as summation, subtraction, multiplication and division.

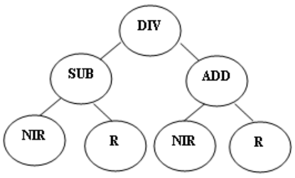

Figure 1. Normalized difference vegetation index as a genetic programming tree solution

Similar to the GA, the GP has parameters and operators such as selection, crossover and mutation, these operators are adapted to work with the tree structure. In addition it has its own parameters such maximum tree depth and size.

Usually a single output value is computed from the operation performed on the values of the attributes based on the expression evolved by the GP. The value computed by the function indicates the class predicted. For binary classification problems; if the output value is greater than a given threshold, the example is assigned to a certain class, otherwise it is assigned to the other one. The most used fitness function in genetic programming is the root mean squared errors between the output values and the desired output of the solution. In our case we used two fitness functions, which permit obtaining well separated groups; the first is based on the using of accuracy of the k-means unsupervised clustering technique. The second is based on the using of the M-statistic.

The M-statistic measures the separability between two classes c_1 and c_2, which are the class of objects of interest and the background (the remaining objects). It can be calculated by normalizing the difference between the means of two classes μ_{c1}– μ_{c2} by the sum of their standard deviations σ_{c1} + σ_{c2}. According to (Kaufman and Remer, 1994) a value of M<1 denotes that the histograms significantly overlap and the ability to discriminate the two classes is poor (figure 3). A value of M>1 denotes that the histogram means are well separated and that the two regions are relatively easy to discriminate.

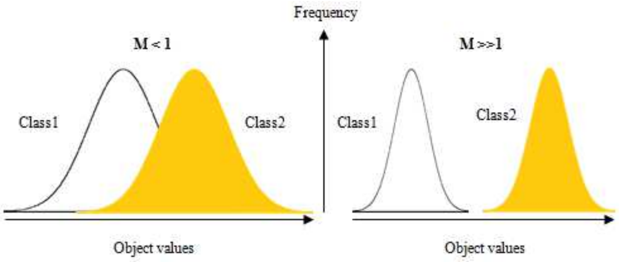

Figure 3. Illustration of the m-statistc group separability measure

2.2.3 Genetic Programming for Rule Induction

Another utilization of genetic programming is the classification rule induction, where instead of evolving mathematical expressions as mentioned previously. GP manipulates IF-THEN classification rules that contain sets of comparators like >, <, IN, and logical operators like AND, OR, NOT. Many algorithms of this kind have been proposed during the last decades. A good survey on the application of genetic programming to classification can be found in (Espejo et al., 2010). In this study we used two algorithms: Discovering interesting classification rules with genetic programming (De Falco et al., 2002) and constrained-syntax genetic programming system for discovering classification rules (Bojarczuk, et al., 2004).

2.2.4 Fuzzy Unordered Rule Induction Algorithm

FURIA is a fuzzy rule-based classification method, which is builds upon the state-of-the-art rule learner algorithm RIPPER.

The sharp boundaries of a crisp rule are replaced by soft boundaries through replacing crisp intervals by fuzzy intervals with trapezoidal membership function (Hühn and Hüllermeier, 2009). A fuzzy interval of that kind is specified by four parameters and will be written $I^F = (\phi^{s,L}, \phi^{c,L}, \phi^{c,U}, \phi^{s,U})$:

$$I^F(v) = \begin{cases} 1, & \text{if} \quad \phi^{c,L} \leq v \leq \phi^{c,U} \\ \frac{v-\phi^{s,L}}{\phi^{c,L}-\phi^{s,L}}, & \text{if} \quad \phi^{s,L} \leq v \leq \phi^{c,L} \\ \frac{\phi^{s,U}-v}{\phi^{s,U}-\phi^{c,U}}, & \text{if} \quad \phi^{c,U} \leq v \leq \phi^{s,U} \\ 0, & \text{otherwise} \end{cases}$$

Where $\phi^{,L}$, and $\phi^{,U}$, are the lower and upper bound of the membership of the fuzzy sets. A fuzzy rule is uniquely characterized by its core [$\phi^{c,L}$, $\phi^{c,U}$] and its support [$\phi^{s,L}$, $\phi^{s,U}$]. It is valid inside the core and invalid outside the support; in-between, the validity drops in a gradual way (Figure 3).

Consider, for example, the rule A<= 5|+, which indicates that if attribute A is smaller or equal to 5, then the class is positive. Here, the rule is valid for A <= 5 and invalid for A > 5. Similarly, a fuzzy rule A ∈ (-∞; -∞; 5; 8)|+ suggests that the rule is completely valid for A <= 5, in valid for A > 8, and partially valid in-between (Hühn and Hüllermeier, 2009).

For details about the FURIA algorithm, reader can refer to (Hühn and Hüllermeier, 2009).

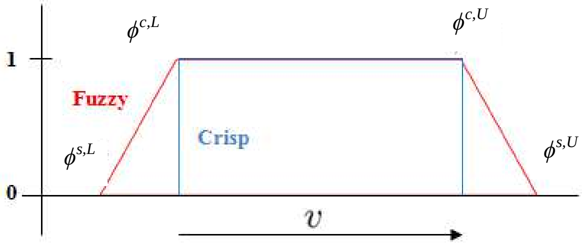

Figure 3. Crisp boundaries vs. Fuzzy boundaries

In contrast to the previously cited algorithms, which generate crisp rules, the FURIA algorithm was applied to the objects dataset to discover fuzzy rules.

3. EXPERIMENTS AND RESULTS

In our experiments, a subset of multi-spectral pansharpened Quickbird image of Oran city, Algeria, is used as the case study image. The image has a ground resolution of 0.6 m and four (04) spectral bands with pixels coded on 11bits.

Firstly, the multi-spectral image is segmented and features are extracted. In this study, numerous combinations of different parameters values were used to yield segmented images, this latter were then checked visually to find the best values for scale level and merge level. Finally the multi-scale segmentation was performed using 30.0 as value for scale parameter and 90.0 for merging threshold, then 38 spectral, texture, spatial, ratio, and color space attributes were computed for each segmented region. Since the selection of training objects influences heavily the quality of the discovered rules. A set of regions of each class were collected and labeled on the basis of field investigation and photo-interpretation. The sample set contains mainly the following classes: buildings, roads, bare soil, trees and grass. These samples are divided to training and test sets. The file containing regions of interest is a XML-based file that was exported later to CSV (comma separated vector) file for further data mining experiments. The training set was used to train different classifiers including genetic algorithm, genetic programming symbolic classification, GP-based classification rule induction (Falco and Bojarczuk algorithms) and Finally the FURIA fuzzy classifier. The test set is applied to evaluate the generalization capability of the final rule set on a separate set containing data examples that not seen during training step.

For the following case studies, we used 500 individuals, 100 generations, and a crossover rate of 90% and a mutation rate of 5% for all the evolutionary algorithms.

Four (04) experiments have been conducted:

1. Extracting vegetation objects. The best extraction rule, the mathematical expression generated by the GP and the rate of correctly classified objects in the test dataset (correctness) are shown in the table 1.
2. Separating tree objects from grass ones. Different results are shown in the table 2.
3. Extracting of built-up areas (buildings and roads. Different results are shown in the table 3.
4. Separating of building objects from roads. Different results are shown in the table 4.

The above experiments have been done according to the hierarchy of figure 4.

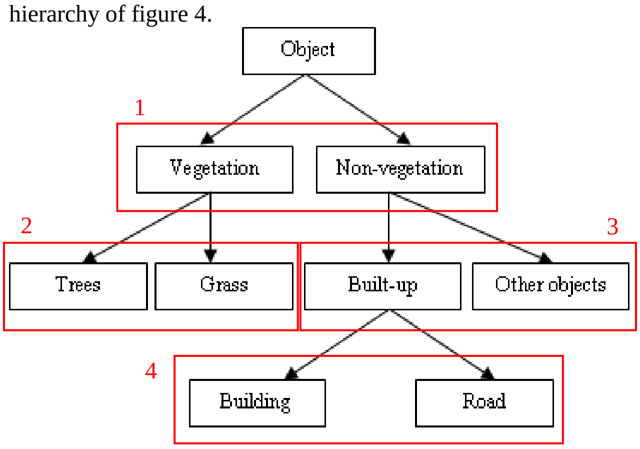

Figure 4. Hierarchy used for the classification of different objects

Classifiers			Rules	Correctness
GA			BANDRATIO IN [0.320 , 0.554]	98.82%
Genetic programming	Symbolic Classification	M-Statistic	Mean(B3)/ Mean(B4) + Mean(B3)/ 2*Mean(B2)	98.31%
		K-means	Mean(B3)+ Mean(B1)- Mean(B4)	98.73%
	Rule induction	Falco	BANDRATIO > 0.255	98.52%
		Bojarczuk	BANDRATIO > 0,2473	98.52%
FURIA			BANDRATIO IN [0.222, 0.223, +∞, +∞]	98.52%

Table 1. Extraction of vegetation results

From the above table, it easy to notice that all classifiers have found the simplest rule to extract vegetation objects, which is based on the thresholding of the normalized difference vegetation index (BANDRATIO), except for the GP-based symbolic classification, where we used only the values of spectral means. The accuracies for the mathematical expressions obtained by the GP are calculated on the test dataset after clustering the output by the k-means algorithm.

The high values of correctly classified instances indicate that separation of vegetation from background is an easy task for all tested classifiers. Even in the case where the bandratio attribute is not available, the GP succeeded in finding simple and easy to memorize formulas for separating between vegetation and non-vegetation segments.

Finding the most adequate attributes and their associated thresholds to extract a particular land-cover type as vegetation can help experts in establishing useful knowledge models (ontologies).

Classifiers			Rules	Correctness
Genetic programming	Symbolic Classification	M-Statistic	AVGBAND_2 - AVGBAND_1 - TX_ENTROPY - AVGBAND_1 + TX_MEAN* TX_ENTROPY	87.0%
		K-means	AVGBAND_1 /[(AVGBAND_2-AVGBAND_1)+ (AVGBAND_2-TX_MEAN+ TX_RANGE)]	87.0%
	Rule induction	Falco	AVGBAND_3 > 203,45	80.37%
		Bojarczuk	AVGBAND_2 > 269,928027	75.70 %
FURIA			AVGBAND_2 IN [-∞, -∞, 261.63, 267.83] AND TX_RANGE IN[26.5982, 26.7438, ∞, ∞]) TX_MEAN IN [-∞, -∞, 180.874, 182.698]) AND (MINBAND_1 IN [138, 139, +∞, +∞]	89.71 %

Table 2. Results of separating tree objects from grass ones

From the table 2. We can conclude that almost algorithms have added new attributes to the rules; among them are texture parameters such TX_MEAN, TX_RANGE and TX_ENTROPY. The obtained rule sets agree with the assumption of domain experts, since trees are considered usually as textured objects. None of the used classifiers added spatial attributes to the rules, which means that this kind of attributes is not useful in this case.

Classifiers			Rules	Correctness
Genetic programming	Symbolic Classification	M-Statistic	AVGBAND_3/2* AVGBAND_1^4	87.13%
		K-means	3*AVGBAND_2- AVGBAND_3-AVGBAND_4	90.08%
	Rule induction	Falco	NOT (TX_RANGE <= 34,94 AND MAXBAND_2 <= 665,157)	78.48%
		Bojarczuk	MAXBAND_1> 320,74 AND AVGBAND_1 > 234,39	82.27%
FURIA			BANDRATIO IN[-∞,-∞, 0.074743, 0.083673]) AND (AVGBAND_1 in [240.166, 241.083, +∞, +∞]) AND (TX_VARIANCE in [257.973, 259.442, +∞,+∞])	93.67%

Table 3. Results of extraction of built-up objects

As it was expected different classifiers have used spectral, bandratio and texture attributes to separate the built-up areas from the remaining objects. The obtained accuracies are acceptable, especially for the GP-based symbolic classification and the FURIA fuzzy classifier.

Classifiers			Rules	Correctness
Genetic programming	Symbolic Classification	M-Statistic	AVGBAND_2^2/ (AVGBAND_1* (AVGBAND_1+ AVGBAND_4))	63.91%
		K-means	(AVGBAND_4+AVGBAND_3)/ (AVGBAND_2- AVGBAND_1)	80.92%
	Rule induction	Falco	COMPACT > 0,147 AND MINBAND_1 > 184,127	88.14 %
		Bojarczuk	AVGBAND_1 > 299,54 AND COMPACT > 0,179737	84.53%
FURIA			COMPACT in [0.167, 0.168, +∞, +∞]) and (MAXBAND_1 in [284, 286, +∞, +∞]	91.23 %

Table 4. Results of separating buildings from roads.

From the table 4. It is obvious that classifiers have succeeded in finding a good way to discriminate between building objects and road objects and that by using the spatial attribute compactness (COMPACT). The best accuracies were obtained by the FURIA algorithm and the FALCO GP-based rule induction. Since the symbolic classification does not use spatial attributes, their accuracies are not as good as the other classifiers.

4. CONCLUSION

Improving classification of very high spatial resolution imagery has become a hot topic in the remote sensing image processing. One of the most innovative classification approach appeared in the era of high resolution earth observation is the object based image classification. This paper has presented a set of data mining techniques to discover classification rules from object-based segmentations. Genetic algorithms and genetic programming are known as flexible and robust algorithms, often capable of solving wide range of optimization problems. Classification rules can be constructed with GA through optimizing intervals of relevant attributes in the object-based space linked with land cover types. Mathematical expressions can be coded as trees and then evolved by GP, which maximize fitness functions as accuracies or classes separability measures. The exploiting of fuzzy rules through the using of algorithms as FURIA is another promising way since it mimics the human behavior in fuzzy situations. These algorithms have been applied to the extraction of classification rules from VHR sub-image of Oran Town (Algeria). The algorithms succeeded in finding comprehensible rules that agree with the domain expert knowledge. Those rules were for the extraction of vegetation areas, the differentiation between grass and trees regions and the distinguishing between roads and buildings. It has been found that the proposed algorithms have comparable accuracies over all experiments, in addition, they provided easy to realize and comprehensible rule set, which can help experts in establishing useful knowledge models.

REFERENCES

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS journal of photogrammetry and remote sensing, 65(1), 2-16.

Bojarczuk, C. C., Lopes, H. S., Freitas, A. A., and Michalkiewicz, E. L., 2004. A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets. Artificial Intelligence in Medicine, 30(1), 27-48.

De Falco, I., Della Cioppa, A., and Tarantino, E., 2002. Discovering interesting classification rules with genetic programming. *Applied Soft Computing*, *1*(4), 257-269.

Easson, G., and Momm, H. G., 2010. Evolutionary computation for remote sensing applications. Geography Compass, 4(3), 172-192.

Espejo, P. G., Ventura, S., and Herrera, F., 2010. A survey on the application of genetic programming to classification. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *40*(2), 121-144.

Freitas, A. A., 2003. A survey of evolutionary algorithms for data mining and knowledge discovery. In Advances in evolutionary computing (pp. 819-845). Springer Berlin Heidelberg.

Hühn, J., and Hüllermeier, E., 2009. FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, 19(3), 293-319.

ITT Visual Information Solutions., 2007. ENVI Feature Extraction Module User's Guide. 48 p.

Kaufman, Y. J., and Remer, L. A., 1994. Detection of forests using mid-IR reflectance: an application for aerosol studies. *Geoscience and Remote Sensing, IEEE Transactions on*, *32*(3), 672-683.

Khelifa, D., & Mimoun, M., 2012. Object-based image analysis and data mining for building ontology of informal urban settlements. In *SPIE Remote Sensing* (pp. 85371I-85371I). International Society for Optics and Photonics

Neubert, M., and Herold, H. 2008. Assessment of remote sensing image segmentation quality. GEOBIA 2008 - Pixels, Objects, Intelligence.

Xiaoying, J., 2007. Segmentation-based image processing system. U.S. Patent Application 11/984,222.