Proceedings of the 8th International Symposium on
Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Shanghai, P. R. China, June 25-27, 2008, pp. 291-297

Scale-Span Classification of Multispectral Images Based on Feature Construction and Decision Trees

Ning Shu [1, 2 +], Liqun Lin [1], Yan Gong [1], Jun Xiao [2] and Fangfang Jin [3]

[1] The School of Remote Sensing Information Engineering, Wuhan University, Wuhan, China

[2] National Lab for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

[3] North Star Power Science and Technology CO., LTD, Hangzhou, China

Abstract. Most of the existing classification methods, based on the homogeneous-region, involve the best segmentation criterion choice. Using the so-called best scale to classify the multi-scale objects defined by human subjectivity, the paper doesn't think it is the best way to correspond with the demand of the scale of human being. So the paper proposes a new scale-span classification method, based on multi-scale homogeneous-region model. The method uses the feature construction to fulfill the construction of scale-span features, and the best scale choice is implicit in the new constructive features, rather than directly carrying on the best scale choice. The experimental result proves the new constructive scale-span features can reduce the dimension of the feature space, and can fully use the longitude information of different scales, thus improve the classification accuracy.

Keywords: feature construction, scale-span, genetic programming, multispectral images, classification, homogeneous-region

1. Introduction

Using the homogeneous-region as the basic image interpretation unit is the hot spot in multispectral image processing in recent years. Generally, the classification based on the homogeneous-region includes the image segmentation [1] and the supervised classification [2]. The purpose of the segmentation is to obtain homogeneous-region; the segmentation result directly affects the classification accuracy [3] [4]. If the segmentation scale is different, the homogeneous-region obtained is different accordingly [4]. Each scale is in the use of one identical objective rule to distinguish different objects, observes the objects from an objective point.

The "Class" is a strong sense of the subjective concept, and in fact implicates the rules that human beings subjectively understand the things. It always happens that the same objects with different perspective might be thought to have different class attributes. It all depends on people how to define. Actually in the sampling process of supervised classification, we are also unable be clear about the scale information, the subjectively defined "Class" may contain different scales, which make the samples themselves contain scale-span information. Then the scale which is the most suitable to distinguish the objects is various.

The paper considers that it might be blind to choose the segmentation scale without previous understanding the scale information of the objects. Existing classification research that are based on homogeneous-region, mostly involves the best segmentation criterion choice. Using the so-called best segmentation criterion to respond the subjective defined multi-criterion objects, obviously, it doesn't satisfy the demand of subjectivity. So in the classification based on the homogeneous-region, it is necessary to

+ Ning Shu. Tel.: +86-27-87161984.
E-mail address: nshuwu@126.com

ISBN: 1-84626-171-6, **ISBN13:** 978-1-84626-171-8

analyze how to meet the demand of different scales, and meanwhile enhance the accuracy of the classification and improve the computer efficiency.

Therefore the paper proposes a new scale-span classification method. It is based on multi-scale homogeneous-region model, fully using the criterion longitudinal information which the model provides. The method uses the feature construction to fulfill the construction of scale-span features, and the best scale choice is implicit in the new constructive features, rather than directly carrying on the best scale choice. The experimental result proves the new constructive scale-span features can reduce the dimension of the feature space, and can fully use the longitude information of different scales, thus improve the classification accuracy.

2. Construction of the multi-scale homogeneous-region model for the multispectral remote sensing images

Multispectral image segmentation aims to obtain homogeneous-region, which are the foundation of the homogeneous-region classification. In accordance with different parameters, most segmentation algorisms can get a variety of different results, which are the multilevel segmentation results. Unfortunately, previous research did not combine the multilevel segmentation results. Therefore we first just need a segmentation way that can change parameters to get multilevel segmentation codes. In order to realize the homogeneous-region analysis, the paper uses the spectral information to do the image segmentation. Segmentation result of each level is an objective realization of the data. Large-scale segmentation reflects the macroscopic characteristic; the small-scale segmentation manifests the microscopic characteristic.

According to the multilevel segmentation codes, if two adjacent points have the same multilevel codes, then we consider these two belong to an identical homogeneous-region. Figure 1 taking three levels of segmentation results as an example, demonstrates using the multilevel segmentation results to construct the multi-scale homogeneous-region model. The bottom D-1 is the large-scale homogeneous-region layer and the top D-2 is the small-scale homogeneous-region layer. And the large-scale homogeneous-region layer is combined by small-scale homogeneous-region layer, so as to ensure the objects under different scales construct the nested hierarchy structures, thus it contributes to extract the homogeneous region attributes between hierarchies.

So through the different grading segmentation scales, the paper establishes a coarse-to-fine multi-scale homogeneous-region model, which expresses the features of the homogeneous-region objects from the macroscopic to the microscopic.

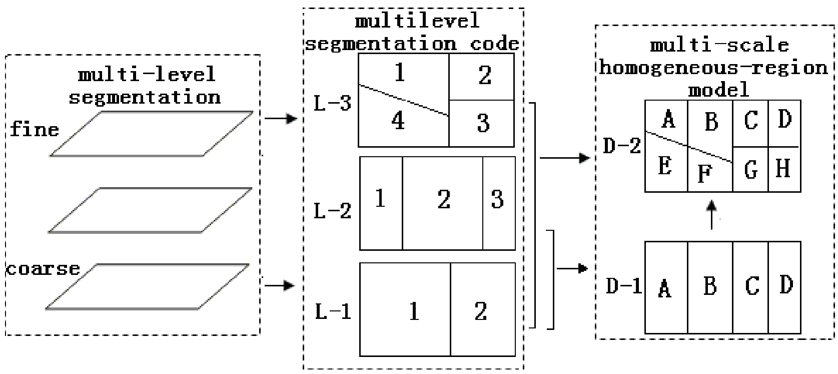


Figure 1. The construction of the multi-scale homogeneous-region model

3. The Scale-span Feature Construction

3.1. the Feature Construction Based on Genetic Programming

The multi-scale homogeneous-region model can make the homogeneous-region of different levels have the inheritance characteristics. Under the multi-scale segmentation, each band has a corresponding multi-scale features. We can use all the features provided by all scales of all bands.

Feature construction [5], or constructive induction concerns the construction of linear or nonlinear combinations of the variables in the original feature space with the aim of discovering highly predictive features, so that the regularities in the data are more easily detected by the classification algorithm, which tends to improve the predictive accuracy.

As the feature construction can construct new attributes out of the continuous attributes of the data set, the paper tries to construct new features out of the feature of all the scale of each band, each new feature of this band is the most useful combination of the features of different scales for the classification. Thus, for the multispectral images, each band will get a new constructive feature. These new features include the scale-span information.

In this paper, in order to fully use the longitudinal information each band provides, the paper uses the genetic programming (GP) to construct the new scale-span feature. GP is a method of searching for this better fitting individual computer program based on Darwinian selection and genetic operations.

We use a standard tree-structure representation for each individual; each individual corresponds to a candidate new attribute. In GP, the genotype is composed of a set of functions and terminal units appropriate to the problem domain. The terminal set consists of all the original attributes plus the constant 1, while the function set consists of the arithmetic operators {+,-,*,/}.

For example, an original feature set includes four features, which are f_1, f_2, f_3 and f_4 respectively. The GP managed to evolve a feature listed in (1), which is a combination of all primitive features in the data set. Figure 2 shows the parse tree representation.

$$F = (f_3 \times f_4) / (f_1 \times f_2) \quad (1)$$

The evolutionary process starting point is the random generation of the initial population. In order to create a new population from the current population, we use three operators, namely reproduction, crossover and mutation. The fitness function used in this work is information gain ratio, which is a well-known attribute-quality measure in the data mining and machine learning literature. Once a GP is over, the solution (constructed feature) returned to the user is the best individual of the last generation.

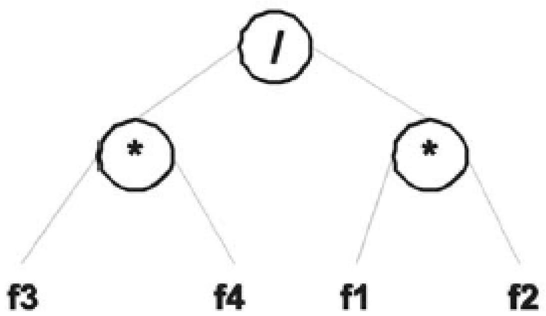


Figure 2: a GP genotype

3.2. The Fitness Measures

As the feature construct can constructs new attributes out of the continuous attributes of the data set, the paper tries to construct new features out of the feature of all the scale of each band, each new feature of this band is the most useful combination of the features of different scales for the classification. Thus, for the multispectral images, each band will get a new constructive feature.

These new features have the following advantages: firstly, each new feature will be a combination of the multi-scale feature of a band, thus include scale-span information; secondly, it reduces the dimension of features, thus reduce the computation time.

Four different fitness measures are used in the GP, each relating to the splitting criteria of one of the decision tree[7] classifiers used in the experiments, namely, Information Gain.

For all four fitness measures, since only numerical attributes are assumed, we adopt the technique used in C5, CART, and CHAID for splitting on a numerical attribute.

Given a set, S of cases with c classes, Entropy(S), given by

$$Entropy(S) = -\sum p_j \times \log_2 p_j$$

Where p_j is the proportion of cased in S belonging to class j, measure the average information needed to identify the class of a case in S.

Likewise, $Entropy_A(S)$ measures the average information needed to identify the class of a case in S when S is partitioned using attribute A:

$$Entropy_A(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i)$$

Where n is the number of partitions caused by attribute A, and $|S_i|$ is the number of cases in S belonging to partition i. The expected reduction in entropy caused by partitioning the examples in accordance with attribute A is designated IG(A):

$$IG(A) = Entropy(S) - Entropy_A(S)$$

Since S is the entire training set, then $Entropy(S)$ is fixed. Thus, the GP needs to evolve an attribute, EA say, that minimizes *IG(EA)*..

4. Experiments and analysis

4.1. Data Explanation

The experiment use TM images (except the 6th band) from Wuhan district, the image size is1024×1024. We first get five different segmentation results, and then use the process shown in Figure 1 to construct the homogeneous-region model of four scales.

According to the multi-scale model obtained above, we conducted the scale-span supervised classification experiment according to nine classes. The purpose of the experiment is to study the feasibility of Scale-Span classification. The 9-class includes: C1-River, C2-lakes, C3-pond (including areas with unusual water quality), C4-old urban areas, C5-new urban areas, C6-vegetation areas in the city, C7-vegetation areas in ridgelines, C8 - farmland, and C9 - bare land.

The features extracted here include the average spectrum of each band and the NDVI average mean; therefore the multi-scale features include four scales and 28 features, consisting the original feature set. The features of homogeneous-regions are represented by Fij($1 \leqslant i \leqslant 7; 1 \leqslant j \leqslant 4$), where F_{ij} indicates the i-th band of the j-th layer, in which j = [1,6] represent the TM1-TM5 and TM7 average spectrum value, j = 7 represents the NDVI feature. Using L1 to L4 to represent the layer from the bottom to the top, the bottom layer L1 is the homogeneous-region of large scale, while the top layer L4 is of small scale.

4.2. Scale-span Feature Construction

Based on the multi-scale homogeneous-region model, we get the spectral and NDVI features of different scales, then we use the GP to construct corresponding features, namely the scale-span features, thus we can get seven scale-span features. Here we use F_i to represent the scale-span feature of i-th band. Table 1 outline the evolved scale-span features, Figs. 2 shows the parse tree representations of F_3.

Table 1 Symbolic Expressions of Some of the Features Evolved Using IG

Band	Expression of the evolved features
1	$F_1 = F_{11} / F_{12}$
2	$F_2 = F_{22} - (F_{21} - F_{22})$
3	$F_3 = (F_{32} + F_{33}) + (F_{34} \times F_{31})$
4	$F_4 = F_{41} \times F_{43}$
5	$F_5 = F_{54} - (F_{52} - F_{53})$
7	$F_6 = (F_{63} / (F_{62} \times (F_{64}/F_{61})) \times (F_{61} + F_{63})$
NDVI	$F_7 = F_{72} / F_{71} - F_{71}$

4.3. Scale-span Classification

According to the training data, we get the classification results and classification rules. Figure 4 is the result of the 9-class, of which subfigure (a) and (b) are the results of the sole scale of L2 and L4 respectively, and subfigure(c) is homogeneous region of scale-span. From the result, as shown in figure 4(a), the two classes C2 and C3 have quite serious confusion, and we can't clearly classify them using the sole scale. While from the figure 4(c) of scale-span classification with the macroscopic characteristic of the low scale as well as the microscopic characteristic of the top level scale, it is quite easy to classify the features in C2 and C3.

The classification rules derived by the decision tree are shown in Figure.5, it can be seen that in scale-span classification, all the new evolved features are involved in classification. Because of these scale-span features, it is easy to classify the classes with serious confusion, such as C2 and C3.

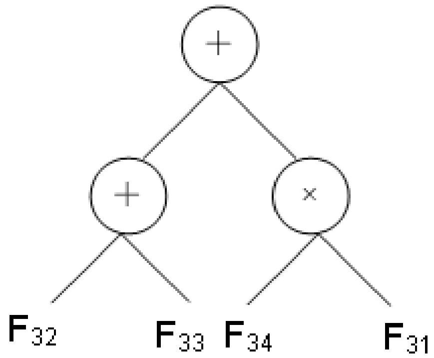
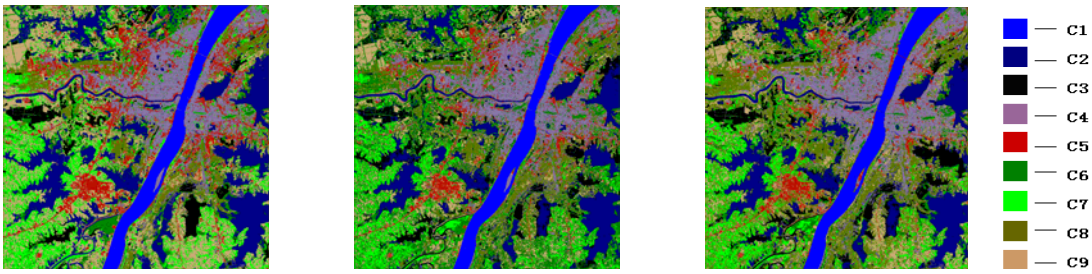


Figure 3 :the parse tree representations of F_3



(a) L2 scale classification result (b) L4 scale classification result (c) scale-span classification result

Figure 4. Classification results

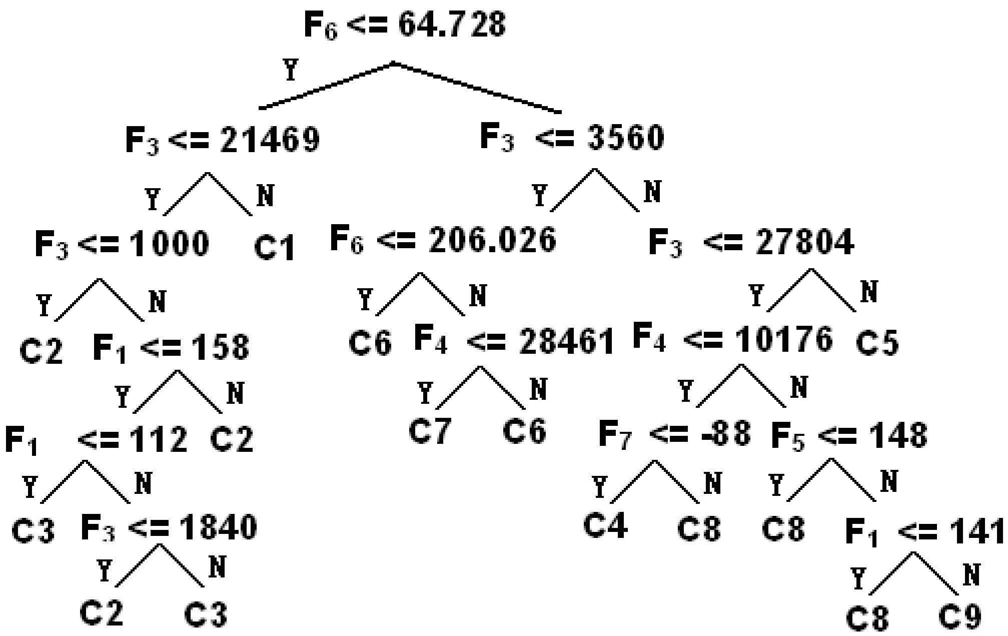


Figure 5: the decision rule

4.4. Data Analysis

According to the test samples, we get the classification accuracy, as shown in table 2, the accuracy of the scale-span classification is 94.8%, respectively increased by 17.4%，4.8%，4.1%和 5.5%, compared with scale 1 to scale 4. So its accuracy is higher than the sole scale classification accuracy. As the class numbers may include different class definition from microscopic or macroscopic perspective, and the scale-span features can completely integrate the macroscopic and the microscopic characteristic, therefore obtain more accurate classification results.

Table 2 Classification accuracy

scale	L1	L2	L3	L4	Scale-span
accuracy	87.4%	90.0%	90.7%	89.3%	94.8%

Table 3 lists the classification accuracy of each class. We can find that, the scale-span has the obvious superiority to the class classification, and the accuracy of each class is much higher than that of sole scale.

Such as in the case of C2, the accuracy of the single scale to this class is low, which accuracy of L1 is 66.7%， of L4 is 80%, while the accuracy of scale-span if 96.7%.

From the data in the table 3, we can also find that, the discrimination of each scale to the classes is various. The single scale can only distinguish some of classes, such as the scale 1 can distinguish the C4, C5 and C6 well, but in the discrimination of C2 and C3 is quite worse. The scale 3 can distinguish the C3 well, but in the discrimination of C2 and C9 is quite worse. Well, the new evolved scale-span features can reach a relatively balanced results in improving the accuracy of various types of classes, thus improve the overall classification accuracy.

Table 3 Accuracy of individual classes

	C1	C2	C3	C4	C5	C6	C7	C8	C9
L1	100%	66.7%	80%	93.3%	100%	90%	90%	86.7%	80%
L2	100%	63.3%	80%	93.3%	100%	86.7%	100%	86.7%	100%
L3	100%	73.3%	93.3%	93.3%	100%	93.3%	100%	80.0%	83.3%
L4	100%	80%	70%	83.3%	100%	96.7%	100%	86.7%	86.7%
Scale-span	100%	96.7%	90.0%	96.7%	100%	100%	96.7%	80.0%	93.3%

From this we can see that it is quite necessary to use the scale-span method in the classification based on homogeneous-region. The construction of the scale-span features can firstly reduce the dimension of feature space, secondly can improve the classification accuracy, and meet the demand of subjectivity.

5. Conclusion

This paper proposes a simple, fast and with high precision classification methods, it's a quite new classification way. Based on the sample training process, it uses feature construction to interpret the subjective scale demand automatically, establishing the relationship between the objective multi-scale homogeneous-region and the subjective classification scales. The paper uses Genetic Programming to remote sensing image processing, and uses it to construct the scale-span features, which can fully use the longitudinal information the homogeneous-region model provides. The result shows this method can improve the classification accuracy and reduce the dimension of feature space.

6. Acknowledgment

Financial supports from key project of Chinese national programs for Fundamental research and development (973 Program)

7. References

[1] Paul S. Hong Lance M. Kaplan Mark J. T. Smith, Hyperspectral Image Segmentation using Filter Banks for Texture Augmentation, *Advances in Techniques for Analysis of Remotely Sensed Data*, 27-28 Oct. 2003:254 – 258.

[2] Antonio J. Plaza, James C. Tilton, Automated Selection of Results in Hierarchical Segmentations of Remotely

Sensed Hyperspectral Images, *IGARSS Proceedings. 2005*,**7**:4946-4949.

[3] Lorenzo Bruzzone. A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images, IEEE *transactions on Geoscience and Remote Sensing*. 2006, **44**(9):2587-2600.

[4] Gong Y., High Spectral Image Homogeneous Area Analysis based on the HDA and the MRF. *Doctor thesis of Wuhan University*,2007,**7**.

[5] Mohammed Muharram and George D. Smith, Evolutionary Constructive Induction, *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17**(11).

[6] Hilan Beansusan, Ibrahim Kuscu, *Constructive induction using genetic programming, International Conference on Machine Learning, Workshop on Evolutionary computing and machine learning*, 1996.

[7] M.A. Friedl and C.E. Brodley, Decision Tree Classification of Land Cover from Remotely Sensed Data, *Remote Sensing of Environment*, 1997, 61(3):399-409.