MULTISPECTRAL LANDSAT IMAGES CLASSIFICATION USING A DATA CLUSTERING ALGORITHM

Yan Wang [a, *], Paul Neville [b], Chandra Bales [b], Mo Jamshidi [a], Stan Morain [b]

[a] Dept. of Electrical and Computer Engineering and Autonomous Control Engineering (ACE) Center, University of New Mexico, Albuquerque, New Mexico, 87131 – yanwang@unm.edu
[b] The Earth Data Analysis Center (EDAC), University of New Mexico, Albuquerque, New Mexico, 87131

KEY WORDS: Multispectral, Image, Classification, Networks, Fuzzy Logic, Algorithms

ABSTRACT:

This paper presents a new application of a data-clustering algorithm in Landsat image classification, which improves on conventional classification methods. Neural networks have been widely used in Landsat image classification because they are unbiased by data distribution. However, they need long training times for the network to get satisfactory classification accuracy. The data-clustering algorithm is based on fuzzy inferences using radial basis functions and clustering in input space. It only passes training data once so it has a short training time. It can also generate fuzzy classification, which is appropriate in the case of mixed, intermediate or complex cover pattern pixels. This algorithm is applied in the land cover classification of Landsat 7 ETM+ over the Rio Rancho area, New Mexico. It is compared with Back-Propagation Neural Network (BPNN) to illustrate its effectiveness and concluded that it can get a better classification using shorter training time.

1. INTRODUCTION

Remotely-sensed imagery classification involves the grouping of image data into a finite number of discrete classes. Conventionally, statistical Maximum Likelihood Classifier (MLC), based on normal distribution assumption, is widely used in remote sensing image classification. However, geographical phenomena do not occur randomly in nature and frequently are not displayed in the image data with a normal distribution. So neural networks with data distribution free have been applied. The neural network classification depends on training data and learning algorithms, which cannot be interpreted by human language or is "a black box". So training data's selection is important for the neural network classification. Normally, the training data sets consist of several thousands of patterns belonging to many (often more than ten) categories and large volumes of data and the neural network structure is complicated to adapt to these patterns. So the neural network training and/or classification time reported are quite long (Heermann, 1992), ranging in some cases from a few hours to a few weeks on a conventional computer. Taking also into account that additional training and classification trials must usually be performed after selecting a particular neural network model, its architecture, and its learning parameters, the need of a methodology for fast neural network training and classification is evident. Vassilas and Charoun (Vassilas, 1999) proposed a methodology based on self-organizing maps and indexing techniques and demonstrated its effectiveness in classifying multispectral satellite images to land-cover categories. In this paper, we propose to use a Radial Basis Function based Clustering (RBFC) algorithm to solve this problem.

In remote sensing images, a pixel might represent a mixture of class covers, within-class variability, or other complex surface cover patterns that cannot be properly described by one class. These may be caused by the ground characteristics of the classes and the image spatial resolution. Since one class cannot describe these pixels, fuzzy classification has been developed. In fuzzy classification, a pixel belongs to a class with a membership degree and the sum of all class degrees is 1. Wang (1990) modified the MLC algorithm with fuzzy mean and fuzzy covariance instead of their conventional counterparts. Foody (1992) embedded the fuzzy concept in all classification stages, including training, classification and evaluation. RBFC is based on fuzzy inferences and the fuzzy rules are generated from the training data. It can also combine human knowledge in it when it is available. The outputs from RBFC are the membership degrees of each class.

In this paper the RBFC is provided and applied the land cover classification of Landsat 7 ETM+ over the Rio Rancho area in New Mexico. Part 2 gives the RBFC algorithm. Part 3 presents a study of land cover classification using RBFC and compares it with Back-Propagation Neural network (BPNN). Part 4 concludes about it.

2. RBFC ALGORITHM

We first describe the mathematical form of the Radial Basis Function (RBF) rulebase, which is identified by the clustering algorithm. We will consider a specific case of a rulebase with n inputs and 1 output. The generalization to m outputs is described in (Berenji, 1993). The inputs to the rulebase are assumed to be normalized to fall within the range [0,1]. Each rule r has the following form, similar to the Takagi-Sugeno-Kang (TSK) rule:
IF s_1 is N (ξ_{r1}, σ_r)... and s_i is N (ξ_{ri}, σ_r)... and s_n is N (ξ_{rn}, σ_r) THEN

* Corresponding author.

$$y_r = \sum_{i=1}^{n} c_{ri} s_i \tag{1}$$

where N (ξ_{ri}, σ_r) is the input membership function with the Gaussian distribution of mean ξ_{ri} and stand deviation σ_r; c_{ri} and ξ_{rI} are tunable coefficients. The weight of rule r for a data point s is determined according to the distance between the vector s and the center of an n-dimensional Gaussian sphere with a mean (ξ_{r1},..., ξ_{rn}) and a standard deviation σ_r, which is the product of input membership values:

$$w_r = \exp\left(-\frac{1}{2\sigma_r^2} \sum_{i=1}^{n} (\xi_{ri} - s_i)^2 \right) = \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma_r^2} (\xi_{ri} - s_i)^2 \right) \tag{2}$$

Normalizing the weights of all rules:

$$\rho_r = \frac{w_r}{\sum_{i=1}^{R} w_i} \tag{3}$$

where R is the number of rules. Finally, the output of the rulebase is given by

$$O = \sum_{r=1}^{R} \rho_r y_r \tag{4}$$

It has been shown that such a configuration can approximate any nonlinear function to any degree of accuracy if the number, the locations, and the variances of Gaussian spheres are allowed to change.

The above approach to identifying an RBF rulebase from data is described in Procedure 1. It begins by creating an RBF rule with the Gaussian center coinciding with the first data point. When the next data point is encountered, the parameters of the first rule are adapted to account for both data points. If the error on the second data point is still too large, then a second rule is created centered at the second data point. The process continues until all data points have been considered. After all the data have been processed, the neurons are pruned to get rid of redundant rules and make knowledge more compact. This may lead to slightly higher error with a reduction in the rulebase. It is a very fast, one-pass algorithm, which also gives very good results as our experiments indicate.

Let s and d be the input and output parts of each data point. For each sample (s, d) do {
INFERENCE:
O = Rulebase output when input s is presented;
Let the index of the nearest Gaussian rule be $i^* = \arg\min_r \{\|s - \xi_r\|\}$;
Let the distance to the nearest rule be $d^* = \{\|s - \xi_{i^*}\|\}$
Let the applicability of the nearest rule be $w^* = \exp(-\|s - \xi_{i^*}\|^2 / 2\sigma_{i^*}^2)$;
Let the error of that rule be $error = \|O - d\|$;
ADAPT_PARAMETERS:
Modify parameters by learning rules:

$$\Delta c_{ri} = \eta(d - O)\rho_r s_i$$

$$\Delta \xi_{ri} = \eta(O - d) y_r \frac{1}{\sigma_r^2} \rho_r (1 - \rho_r)(\xi_{ri} - s_i)$$

$$\Delta \sigma_r = -\eta(O - d) y_r \frac{1}{\sigma_r^3} \rho_r (1 - \rho_r) \|\xi - s\|^2$$

ADD_RULE:
If ($w^* < \delta$){
add neuron at s with spread $d^* / \sqrt{2 \ln 2} - \sigma_{i^*}$;
INFER and ADAP_PARAMETERS;
} else if (error > ε){
add neuron at s with spread σ_{min}
INFER and ADAP_PARAMETERS;
}
}
PRUNE_RULES:
For each pair of remaining neurons n_a and n_b ($a < b$) do {
Let θ_{jab} = angle (hyperlane$_{aj}$, hyperplane$_{bj}$);
Let $d_{ab} = \|\xi_a - \xi_b\|$;
If (max$_j$ $\theta_{jab} < \omega$){
if ($\sigma_a > d_{ab}$){winner: = a; loser: = b}
else if ($\sigma_a < d_{ab}$){winner: = b; loser: = a}
else consider next pair;
move winner towards loser in proportion $\sigma_{winner}^n : \sigma_{loser}^n$;
expand winner's radius to include loser's radius;
delete loser neuron;
}
}
EXTRACT_RULES from neurons;

Procedure 1. Algorithm for extracting radial basis function rules from data

To verify the effectiveness of the RBFC algorithm in the Landsat image classification, in the following section, we will compare it with the widely used multilayer neural network, Back Propagation Neural Network (BPNN) (Duda, 2001; Heermann, 1992).

3. A STUDY OF LAND COVER CLASSIFICATION

3.1 Landsat 7 ETM+ Data Set Over Rio Rancho

Landsat 7 carries the Enhanced Thematic Mapper Plus (ETM+) instrument—a nadir-viewing, multispectral scanning radiometer, and provides image data for the Earth's surface via eight spectral bands (NASA, 2000; USGS Landsat 7, 2000). The bands are for the visible and near infrared (VNIR), the mid-infrared (Mid_IR), and the thermal infrared (TIR) regions of the electromagnetic spectrum, as well as the panchromatic region. Table 1 lists the ETM+ Bands, spectral ranges, and nominal ground resolution. Data are quantized at 8 bits. The size of the image for one band is 744 lines × 1014 pixels, which is shown in Figure 1 over the Rio Rancho area with 3 bands

TM1, TM4, TM6 in blue, green and red, respectively. We will use these three bands as the inputs to the RBFC and BPNN because they represent most discrimination among classes. This site mainly contains eight types of land cover, which are water, urban imperious, irrigated vegetation, barren, bosque, shrubland, natural grassland and juniper. In the urban area, buildings are blocked by streets and sometimes covered with vegetation. So the degree of class mixture in the urban area is high. Besides, shrubland, natural grassland and juniper are highly mixed too. These bring up the difficulty in the land cover classification. For comparison both RBFC and BPNN are used to generate the land cover map of Figure 1.

Band Number	Spectral Range (μm)	Ground Resolution (m)
TM1 (Vis-Blue)	0.450 - 0.515	30
TM2 (Vis-Green)	0.525 - 0.605	30
TM3 (Vis-Red)	0.630 - 0.690	30
TM4 (NIR)	0.750 - 0.900	30
TM5 (Mid-IR)	1.550 - 1.750	30
TM6 (TIR)	10.40 - 12.50	60
TM7 (Mid-IR)	2.090 - 2.350	30
TM8 (Pan)	0.520 - 0.900	15

Table 1. Landsat 7 ETM+ bands, spectral ranges, and ground resolutions

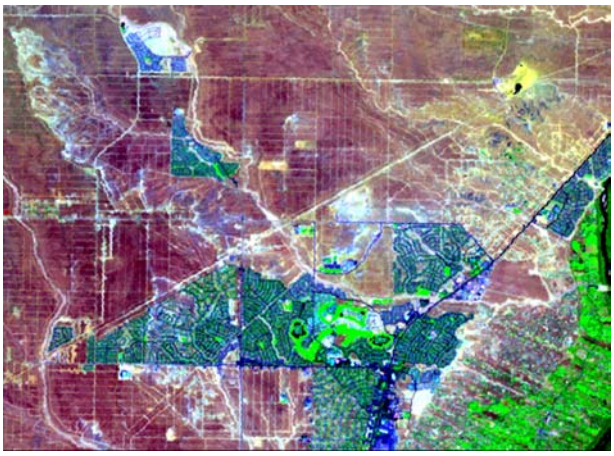

Figure 1. ETM+ image with 3 bandsTM1, TM4, TM6 displayed in blue, green, and red, respectively

3.2 BPNN classification

There are three input nodes, eight output nodes and one hidden layer with ten nodes for BPNN. First, each band is normalized to be in [0, 1]. The output node representing a class is defined as 1 (unity) if the input data point belongs to the class, otherwise 0. The training of BPNN adaptively adjusts the learning rate and the momentum [7]. It runs 10000 epochs and all the training data are computed once for one epoch, so each data point is passed 10000 times. The classification result is shown in Figure 2. As shown in Table 2, the classification accuracy of urban impervious, shrubland, natural grassland and juniper are low because the mixed pixels in these classes are not distinguished well.

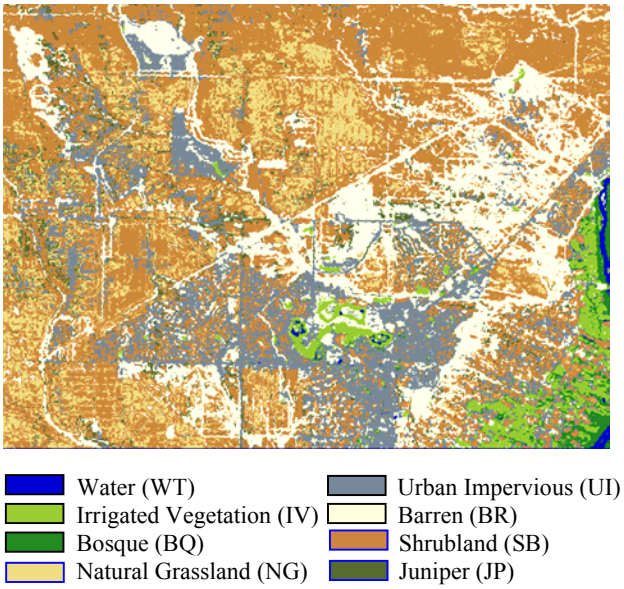

Figure 2. BPNN classification result

3.3 RBFC Classification

The input and output nodes of RBFC are defined the same as those of BPNN. The number of nodes in RBFC is determined in the training process and it uses 1000 nodes (or rules) to adapt to the training data. Each data point is only passed once in the training so it reduces the training time a lot. The order of training data is first randomized to reduce the training error. The classification result is shown in Figure 3 and the accuracy matrix of RBFC is shown in Table 3. The classification accuracy of urban impervious is improved from 88.84 percent to 95.31 percent. At the same time, the classification accuracy of shurbland is reduced from 86.51 percent to 71.43 percent and the classification accuracy of barren is reduced from 97.29 percent to 94.86 percent. However, the classification accuracy of natural grassland is increased from 28.36 percent to 31.34 percent and the classification accuracy of juniper is increased substantially from 36 percent to 55.43 percent. So the overall accuracy is slightly increased a little from 88.46 percent to 88.64 percent.

One of the rules in RBFC is: IF s_1 is N (0.93, 0.19), s_2 is N (0.95, 0.19) and s_3 is N (0.58, 0.19) THEN $y_1 = 0.00058s_1 + 0.00042s_2 - 0.00019s_3 + 0.00025$,..., $y_9 = 0.00280s_1 - 0.00556s_2 + 0.00678s_3 + 0.00730$. The input space is distributed with 1000 clusters and there are more rules where the outputs change greatly. The outputs (O_1,..., O_m) from RBFC are satisfied with $0< O_i <1$ and $\sum_{i=1}^{m} O_i = 1$ after the RBFC is trained well. RBFC gives the fuzzy classification result, and it is constructed by interpretable fuzzy rules.

4. CONCLUSION

A radial basis function based clustering method is used in mulispectral image land cover classification. It improves the training process tremendously because the training data is passed once to the algorithm. Its effectiveness is demonstrated by the study of Landsat 7 ETM+ image classification. It can

Actual Class	Predicted Class								Accuracy (%)
	WT	UI	IV	BR	BQ	SB	NG	JP	
WT	223	0	1	0	2	0	0	0	98.67
UI	0	852	1	55	0	43	3	5	88.84
IV	0	1	522	2	1	0	0	0	99.24
BR	0	12	0	1402	0	3	3	21	97.29
BQ	0	0	0	0	81	0	0	0	100
SB	0	19	0	2	0	327	25	5	86.51
NG	0	4	0	36	0	99	57	5	28.36
JP	0	8	0	45	0	47	12	63	36.00
	Average accuracy (%) = 79.36 Overall Accuracy (%) = 88.46								

Table 2. Classification matrix for the study area by using BPNN

Actual Class	Predicted Class								Accuracy (%)
	WT	UI	IV	BR	BQ	SB	NG	JP	
WT	222	0	1	0	3	0	0	0	98.23
UI	0	914	0	29	0	8	0	8	95.31
IV	0	6	520	0	0	0	0	0	98.86
BR	0	39	0	1367	0	4	0	31	94.86
BQ	0	0	0	0	81	0	0	0	100
SB	0	63	0	4	0	270	6	35	71.43
NG	0	8	0	28	0	86	63	16	31.34
JP	0	11	0	43	0	21	3	97	55.43
	Average accuracy (%) = 80.68 Overall Accuracy (%) = 88.64								

Table 3. Classification matrix for the study area by using RBFC

slightly improve the classification accuracy compared with BPNN. It can provide fuzzy classification results, more appropriate in the case of mixed, intermediate, or complex cover pattern pixels. The structure of this algorithm is composed of a limited number of fuzzy rules, which are interpretable and can be modified by human knowledge.

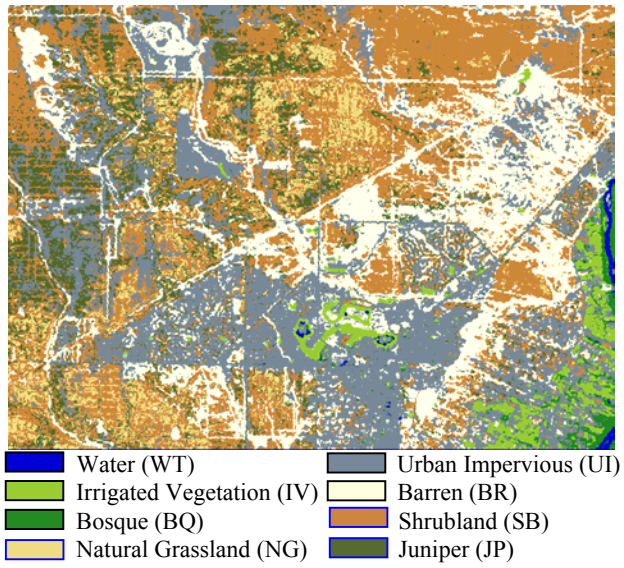

Figure 3. RBFC classification result

5. ACKOWELDGEMENT

The authors wish to thank the Earth Data Analysis Center (EDAC) at UNM for the Landsat data set and grateful help. We also thank the support and discussion of students and staff of the Autonomous Control Engineering (ACE) Center at UNM.

6. REFERENCES

Berenji, H. R. and Khedkar, P. S., 1993, "Clustering in Product Space for Fuzzy Inference", *IEEE Fuzzy*, pp. 1402-1407.

Duda, R. O., Hart, P. E. and Stork, D. G., 2000, *Pattern Classification*, Jone Wiley & Sons (Asia) Pte. Ltd., New York, 2001.

Foody, G.M., 1992, "A fuzzy sets approach to the representation of vegetation continua from remotely sensed data: An example from lowland heath", *Photogramm. Eng. Remote* Sens., 58(2), pp. 443–451.

Heermann, P.D. and Khazenic, Nahid, 1992, " Classification of Multispectral Remote Sensing Data using a Back-Propagation Neural Network", *IEEE Trans. Geoscience and Remote Sensing*, 30(1), pp. 81-88.

Vassilas, N. and Charou, E., 1999, "A New Methodology for Efficient Classification of Multispectral Satellite Images Using Neural Network Techniques", *Neural Processing Letters*, 9, pp. 35-43.

Wang, F., 1990, "Improving Remote Sensing Image Analysis Through Fuzzy Information representation," *Photogramm. Eng. Remote* Sens., 56(8), pp. 1163–1169.