Can. J. Remote Sensing, Vol. 33, No. 3, pp. 203–213, 2007

A genetic algorithm (GA) based automated classifier for remote sensing imagery

Ming-Der Yang

Abstract. Conventional unsupervised classification divides all pixels within an image into corresponding classes based on the distance between pixels and the cluster centres. The number of classes must be selected a priori but is seldom ascertainable with little information. To analyze a large dataset, such as a remote sensing dataset, requires an automatic unsupervised classifier which needs no human effort during the process of image clustering. A genetic algorithm (GA) is adopted to search the cluster centres and choose a suitable cluster number for digital images to overcome the disadvantages of the conventional unsupervised classifier. The GA-based automated classifier was executed on several test images for validity and SPOT satellite imagery for practical application. The satellite images classified by the GA-based classifier and iterative self-organizing data analysis technique (ISODATA) were compared with a classified result through a supervised classification. According to the estimation of classification accuracy by error matrices and $\hat{K}$ statistic, the GA-based classifier performed better than the unsupervised ISODATA and as good as a supervised classifier, even without manipulation by an analyst. A modified GA-based classifier using maximum likelihood (represented by the z score) as a clustering criterion was also proposed and proven to be capable of performing automatically as well as a supervised classifier.

Résumé. La classification non dirigée conventionnelle divise tous les pixels à l'intérieur de l'image en classes correspondantes sur la base de la distance entre les pixels et les centres des regroupements. Le nombre de classes doit être sélectionné a priori, mais ce nombre est difficile à évaluer lorsque l'on dispose de peu d'information. Pour analyser un gros ensemble de données comme c'est le cas en télédétection, il est nécessaire d'avoir un classifieur automatique non dirigé qui ne requiert aucune intervention humaine durant le processus d'analyse des regroupements de l'image. L'algorithme génétique (AG) est adopté pour rechercher les centres des regroupements ainsi qu'un nombre satisfaisant de regroupements pour que les images numériques puissent s'affranchir des inconvénients du classifieur non dirigé conventionnel. Le classifieur automatisé basé sur l'AG a été utilisé sur plusieurs images tests pour la validation et sur des images de SPOT pour une application plus pratique. Les images satellitaires classifiées au moyen du classifieur AG et d'ISODATA (« iterative self organizing data analysis technique ») ont été comparées avec un résultat de classification par le biais d'une classification dirigée. L'estimation de la précision de classification utilisant les matrices d'erreur et les statistiques K a montré que le classifieur basé sur l'AG affiche une meilleure performance que l'ISODATA non dirigé et une aussi bonne performance que le classifieur dirigé même sans manipulation par l'analyste. Un classifieur modifié basé sur l'AG utilisant le maximum de vraisemblance (représenté par la note z) comme critère de regroupement a aussi été proposé et a montré sa capacité d'agir également de façon automatique comme classifieur dirigé.
[Traduit par la Rédaction]

Introduction

Image classification, including supervised and unsupervised classification, is a major analytical procedure in digital image processing (Lillesand and Kiefer, 2000). Supervised classification procedures require the analyst to provide training areas, which are groups of pixels with known identities, to assemble groups of similar pixels into a proper class (Avery and Berlin, 1992). In comparison, unsupervised classification divides all pixels within an image into corresponding classes pixel by pixel and proceeds with fewer interactions with the analyst. Unsupervised clustering techniques are broadly used for exploratory data analysis. Unsupervised classification on remote sensing imagery can be defined as the identification of natural groups within multidimensional data and is an essential step in automatic pattern recognition. A typical unsupervised classification requires a specific number of classes based on the analyst's knowledge of the scene. However, the analyst seldom has sufficient information to decide on a suitable cluster number. In many cases, the given cluster number results in an improper classification, and new runs have to be performed from scratch or several clusters with greater similarity have to be merged based on the experience of the analyst.

Recently, clustering techniques have been applied to vast digital datasets, such as (*i*) medical images for diagnosing tumors as benign or malignant in mammographs (Guliato et al., 2003a; 2003b), segmenting bone and soft tissue in radiographs (Pakin et al., 2003), and discriminating myocardial heart disease from echocardiographs (Tsai et al., 2004); and (*ii*) remote sensing images for land use analysis (Miller et al., 1995; Mohanty and Majumdar, 1996; Bandyopadhyay and

Received 22 September 2005. Accepted 19 April 2007. Published on the *Canadian Journal of Remote Sensing* Web site at http://pubs.nrc-cnrc.gc.ca/cjrs on 19 July 2007.

M.-D. Yang. Department of Civil Engineering, National Chung Hsing University, 250 Kuo-Kuang Road, Taichung, Taiwan, Republic of China (e-mail: mdyang@dragon.nchu.edu.tw).

Maulik, 2002; Maulik and Bandyopadhyay, 2003), agriculture monitoring (Rydberg and Borgefors, 2001; Murthy et al., 2003), and natural hazard investigation and management (Ostir et al., 2003; van der Sande et al., 2003; Yang et al., 2004; 2007). For civil and environmental engineers, clustering techniques for practical applications are expected to detect the earth terrain on remote sensing images automatically. Spectral properties of specific informational classes of remote sensing imagery change temporally, so the relationships between informational classes and spectral classes are not always constant, and relationships defined for one image cannot be extended to others. In addition, the analyst has very limited knowledge about the menu of classes and their specific identities in most cases. With an unknown cluster number a priori, the computational process and clustering accuracy of unsupervised classification remain to be improved.

The aim of this research is to develop a repeatable, accurate, and time-effective method to classify remote sensing imagery automatically. A genetic algorithm (GA) based classifier was established for solving a multidimensional unsupervised classification problem to result in a best partition without prior knowledge of the clustering number. The GA classifier was encoded and tested on two artificial datasets with known cluster numbers and cluster centres and a real image with an unknown cluster number and cluster centres. The GA classifier was then applied to a satellite image to identify a landslide area in central Taiwan.

Methodology

Classical clustering algorithm

Clustering techniques can be broadly divided into two categories, namely hierarchical and nonhierarchical (Murthy and Chowdhury, 1996; Tseng and Yang, 2001). *K* means (or *C* means), which is one of the most popular nonhierarchical algorithms, optimizes an objective function that is the minimum of the sum of squared Euclidean distances between patterns and cluster centres (Murthy and Chowdhury, 1996; Bischof et al., 1999; Nascimento et al., 2003). A systematical solution to an optimal clustering decision is to apply a given clustering algorithm for a range of *K* values and then evaluate the validity of the resulting partition in each case (Dave and Krishnapuram, 1997). In other words, the clustering problem is to group a set of data objects into *K* desired clusters by optimizing an objective function of high intracluster similarity and low intercluster similarity (Mitra, 2004). However, the clustering must be executed for every value of *K* over a specific range and requires a large and costly computation. Based on *K* means, the iterative self-organizing data analysis technique (ISODATA) algorithm has been developed and is the most popular method of unsupervised classification easily found in the public domain (Pierce et al., 1998). Thus, an ISODATA built-in image processing software, ERDAS, was used to run the cases for comparison in this research. The ISODATA algorithm starts with the analyst specifying a number of categories and a classification criterion. The classifier then calculates and assigns each pixel individually with a set of arbitrarily selected pixels as cluster centres over the entire scene. Next, new centres for each class are found and the entire scene is classified again. The preceding steps are repeated until there is no significant change detected in locations of class centres. The initial number of classes is commonly set larger than the possible actual number of classes in the field. Consequently, the iteration of merging classes is performed based on the mean and covariance matrix, which requires a large computation for large objects. Thus, innovative computational approaches were developed to efficiently search the cluster centres, such as GAs and neural networks (NNs) (Alippi and Cucchiara, 1992; Miller et al., 1995; Oin and Suganthan, 2004; Tsai et al., 2004). Most neural network algorithms include a training procedure that is an obstruction to turn into a completely automatic classifier. GAs are particularly suitable for solving complicated optimization problems in situations where uncertainty and imprecision exist (Alippi and Cucchiara, 1992). GAs have been used in a wide variety of optimization problems, specifically in classifying digital datasets (Alippi and Cucchiara, 1992; Zhang and Wang, 1994; Ross, 1995; Diederich and Fortuner, 1999; Tseng and Yang, 2001; Bandyopadhyay and Maulik, 2002; Maulik and Bandyopadhyay, 2003; Garai and Chaudhuri, 2004; Yang and Su, 2006). However, human participation in several steps during classical clustering remains a hurdle to changing the unsupervised classification to an automated technique. By employing a proper clustering index as fitness, a GA with a length-variable chromosome can determine the most suitable number of clusters and the most proper cluster centres at a lower computation cost.

Automated GA-based classifier

There are several steps to establish a GA classifier for automated clustering, including encoding chromosome strings, defining a fitness function, and executing genetic operations (Ross, 1995).

Encoding chromosome

In GA applications, the parameters of the searched space are encoded in the form of strings, so-called chromosomes, representing a solution of problems and being encoded by a binary number, an integer, or a real number. Without assigning the number of classes a priori, a variable string length is designed (Maulik and Bandyopadhyay, 2003; Yang and Yang, 2004). In this research, a chromosome is encoded by positive real numbers in which "0" represents a nonexistent cluster. The value of *K* (valid clusters) is randomly assumed in the range $[K_{min}, K_{max}]$, where K_{min} is usually assigned a value of 2 unless special cases are considered, and K_{max} is the length of a chromosome. In a chromosome, each individual gene represents either a cluster centre or a nonexistent cluster. For each chromosome i in the population ($i = 1, 2, \ldots N$, where N is the size of the population), K_i points are chosen randomly from

the dataset and then are randomly allocated in the chromosome. Take the following case as an example. Assuming an image including three bands as a classified digital dataset, N pixels for each layer, $K_{\min} = 2$, $K_{\max} = 8$, and $K_i = 5$ for the chromosome i, let the five cluster centres be as follows:

110	88	246	150	78	226	11	104	8	50	100	114	227	250	192

Randomly, the classification centres can be encoded into a chromosome as follows:

0	110	88	246	150	78	226	0	11	104	8	50	100	114	0	227	250	192

Fitness definition

Before a GA is operated, an objective function needs to be defined to measure the fitness of each chromosome. In this research, the main objective is to find the best clustering for remote sensing imagery. The population evolves over generations in an attempt to maximize the fitness, which is considered the clustering validity in this research and assigns an adaptability degree to each chromosome in the population. Several clustering validity indices were developed to determine optimal clustering, such as the separation index (SI), the Daviers–Bouldin (DB) index, the Xie–Beni (XB) index, Hubert statistics, and the Dunn index (DI) in which the DB index has both a statistical and a geometric rationale (Ross, 1995; Bezdak and Pal, 1998; Groenen and Jajuga, 2001; Bandyopadhyay and Maulik, 2002; Maulik and Bandyopadhyay, 2003). The minimum description length (MDL) was also used to determine the optimal number of clusters (Bischof et al., 1999; Oin and Suganthan, 2004). The GA classifier adopts the DB index to represent the fitness of a chromosome because of its suitability for remote sensing imagery. The DB index can be calculated as follows (Xie and Beni, 1991; Bezdak and Pal, 1998; Swanepoel, 1999; Groenen and Jajuga, 2001; Martini and Schobel, 2001; Yang and Wu, 2001):

$$u_{ij} = \begin{cases} 1 & \min\limits_{j=1 \text{ to } K} \left\| x_i - v_j \right\| \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$v_j = \frac{\sum\limits_{\text{all}} u_{ij} x_i}{\sum\limits_{\text{all}} u_{ij}} = \frac{\sum\limits_{\text{all}} u_{ij} x_i}{C_j} \tag{2}$$

$$S_j = \left(\frac{\sum\limits_{x_i \subset C_j} \left\| x_i - v_j \right\|}{\left| C_j \right|} \right)^{1/2} \tag{3}$$

$$d_{kj} = \| v_k - v_j \| \tag{4}$$

$$\text{DB} = \frac{1}{K} \sum_{j=1}^{K} \max_{j \neq k} \left\{ \frac{S_k + S_j}{d_{kj}} \right\} \tag{5}$$

where u_{ij} is the membership of each x_i belonging to the jth cluster; x_i is any pixel in the image ($1 \le i \le N$, where N is the total number of pixels in the image); v_j is the centre of the jth cluster ($1 \le i \le K$, where K is the total number of clusters); S_j is the standard deviation of the jth cluster; C_j is the dataset of the jth cluster; $|C_j|$ is the pixel number of the jth cluster; d_{kj} is the Euclidean distance between the kth and jth centres; v_k is the other centres of the clusters ($1 \le k \le K$; and $k \neq j$); S_k is the standard deviation of the kth cluster; and $\|*\|$ denotes the norm for Euclidean distance calculation.

The DB index is defined as the averaged optimal ratio of the intracluster scatter over the intercluster separation. Thus, the fitness function for chromosome j is defined as 1/DBj. The maximization of the fitness function ensures a minimum DB value, which means the optimal clustering with the smallest intracluster scatter and the largest intercluster separation. Similarly, the ratio of the difference between class centres over the sum of their standard deviations is called the normalized difference, which is a typical measurement of the distinctiveness between classes generally adopted in remote sensing classification.

Genetic operations

In general, a GA is composed of three operators, namely reproduction, crossover, and mutation. Reproduction calculates a survival probability of each chromosome which is a criterion to reproduce better chromosomes for the next generations. The operation follows Darwinism: natural selection and survival of the fittest. Crossover is a swapping process to create new chromosomes between the reproduced chromosomes. To avoid sticking to a local optimal, mutation is assigned to explore the possible optimal in all the space. The mutation probability is usually set smaller than the crossover and controls the percentage to introduce new genes for trial. If the mutation probability is too low, some useful genes are not discovered; on the contrary, if it is too high, there will be severely random perturbation (Gen and Cheng, 1997). These operations are

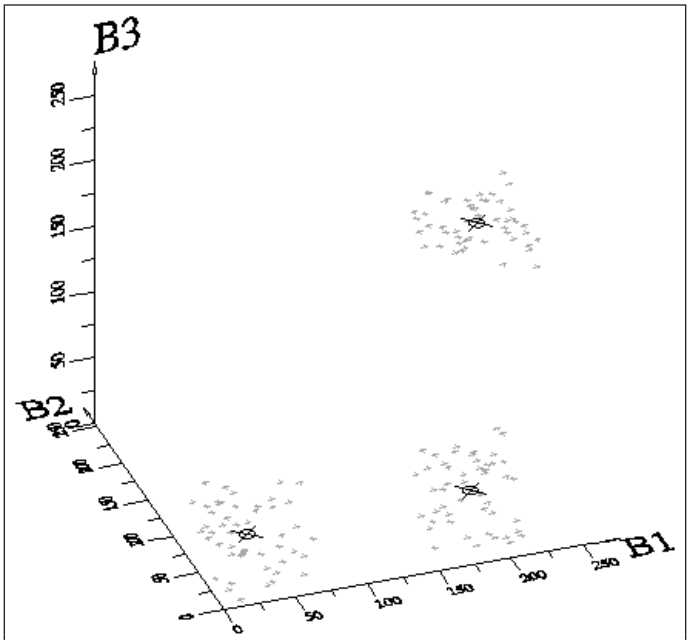

Figure 1. Distribution of points and cluster centres for Data_1.

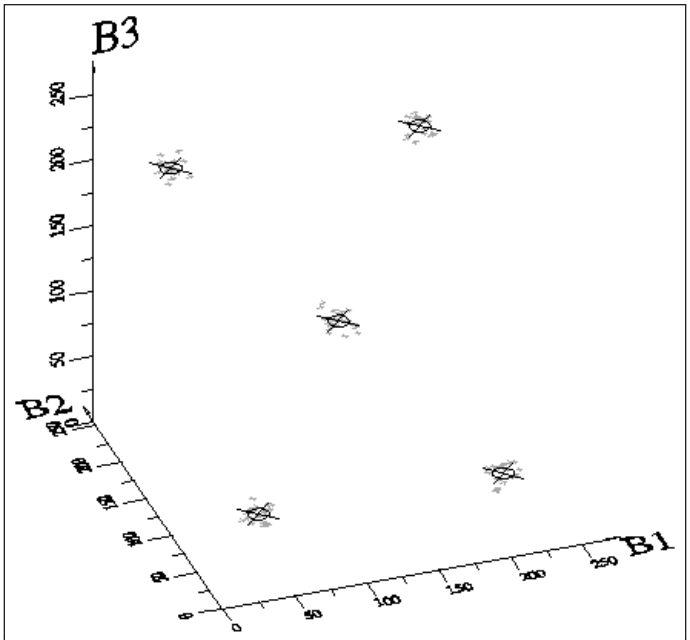

Figure 2. Distribution of points and cluster centres for Data_2.

repeated until the terminal criterion is satisfied and the best chromosome is found. In following GA classification, population size is equal to 80 chromosomes, and the maximum iteration is 200 generations. After the calculation of fitness for each chromosome in the population, the reproduction operator is implemented by stochastic universal sampling (SUS), which is a single-phase sampling algorithm with minimum spread and zero bias, instead of the single selection pointer employed in roulette wheel methods (Chipperfield et al., 1994). Regarding crossover operation, 80% antecedents are swapped using a

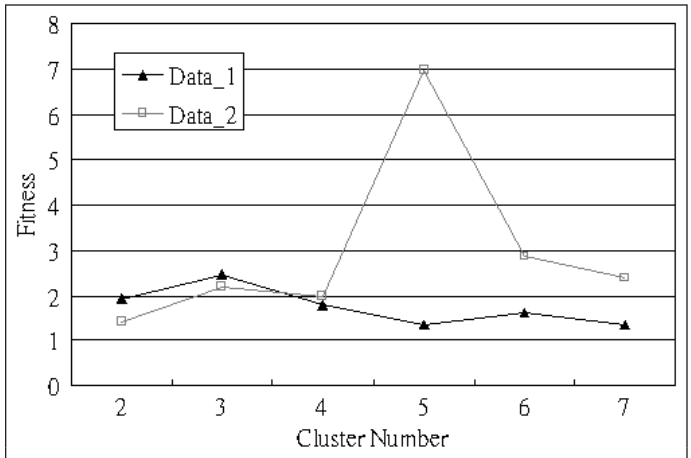

Figure 3. Superior fitness versus cluster numbers for Data_1 and Data_2.

uniform probability. The mutation rate is 1, which is comparatively low because the newly introduced chromosomes generate considerable disturbance. All GA operations were programmed in MATLAB®, version 6.5 (The MathWorks, Inc., 2002), and were conducted using a personal computer with a Pentium IV processor at 1.4G Hz.

Validation of the GA-based classifier

Two artificial datasets, Data_1 and Data_2 shown in **Figures 1** and **2**, were designed as three-dimensional (3D) point-basis datasets for validation of the GA classifier. Data_1 is an irregularly tiered dataset with three clusters, and Data_2 is characterized as a dataset with five spherical clusters. Both datasets have distinctive features, especially Data_2. Clustering centres of Data_1 and Data_2 were known a priori as the comparable data. **Figure 3** shows that the superior fitness varies with the assigned clustering numbers. Three clusters for Data_1 and five clusters for Data_2 found by the GA classifier are identical to the original design. In **Figure 3**, the maximum fitness of Data_1 is lower than that of Data_2, illustrating that Data_1 less distinguishable than Data_2 as shown in **Figures 1** and **2**.

Table 1 lists the sizes of the original datasets, the actual and classified numbers of clusters, and the centres of the clusters through the classification of ISODATA and the GA classifier. The GA classifier can also accurately identify the cluster centres. Apparently, the GA classifier is able to classify both datasets quite accurately, which assures that the validation of the GA classifier is affirmative.

Application of the GA-based classifier to image classification

After being examined using the artificial data, the GA classifier was applied to real imagery data, namely a picture of a dolphin (Data_3, see **Figure 4**) and a satellite image (Data_4, see **Figure 5**).

Table 1. Classification results for Data_1 and Data_2.

Type of dataset	Total no. of points	No. of actual clusters	No. of classified clusters	Cluster centres		
				Designed	ISODATA	GA classifier
Data_1						
Three irregular clusters	150	3	3	{(30.30, 39.32, 29.72) (183.66, 34.42, 35.64) (187.64, 33.62, 239.24)}	{(30.30, 39.32, 29.72) (183.66, 34.42, 35.64) (187.64, 33.62, 239.24)}	{(30.30, 39.32, 29.72) (183.66, 34.42, 35.64) (187.64, 33.62, 239.24)}
Data_2						
Five spherical clusters	100	5	5	{(39.90, 43.00, 40.25) (38.55, 212.10, 210.00) (211.15, 211.25, 209.00) (209.45, 41.20, 38.85) (124.35, 125.65, 125.05)}	{(39.90, 43.00, 40.25) (38.55, 212.10, 210.00) (211.15, 211.25, 209.00) (209.45, 41.20, 38.85) (124.35, 125.65, 125.05)}	{(39.90, 43.00, 40.25) (38.55, 212.10, 210.00) (211.15, 211.25, 209.00) (209.45, 41.20, 38.85) (124.35, 125.65, 125.05)}

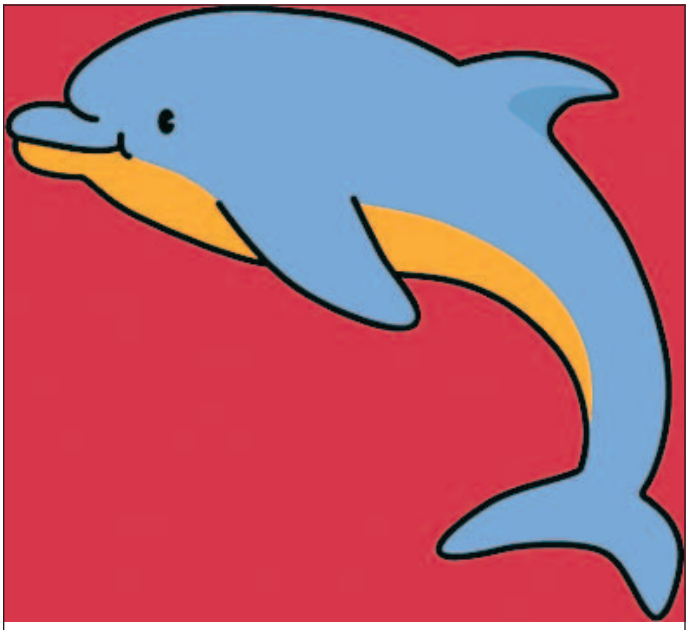

Figure 4. Original dolphin picture (Data_3).

Characteristics of real images

Real images are usually more complicated and ambiguous than point-basis datasets and are also quite large. For remote sensing imagery, in particular, the clusters are not usually as distinguishable as the artificial data because pixels of intermediate values tend to fill in the gaps between groups (Avery and Berlin, 1992). In addition, the influencing factors of image brightness, such as illumination, shadowing, and mix pixels, may produce extra variations in cluster partitioning. However, remote sensing imagery is composed of several spectral channels that provide multidimensional information to identify various clusters.

Four clusters can be visually identified in Data_3, which originally was a colorful picture as shown in **Figure 4**. **Figure 6** displays the pixel distribution of the dolphin picture in a 3D red–green–blue (RGB) space (where R is band B1, G is band B2, and B is band B3).

A multispectral SPOT4 imag (with a path of K299 and a row of J303) has about 9 620 000 pixels for each band and 28 860 000 pixels in total. For a clear display, only 15 752 pixels (rectangular subset in **Figure 5**) as Data_4 are shown and discussed herein. A SPOT4 XS image of the Tsao-Ling region in Taiwan was taken on 11 February 2000 after the Chi-Chi earthquake, the most serious disaster in Taiwan during the last century. The original image is a false-color image composed of three bands, namely green (wavelength 0.50–0.59 μm, band 1 or B1), red (wavelength 0.61–0.68 μm, band 2 or B2), and near infrared (wavelength 0.79–0.89 μm, band 3 or B3). **Figure 7** shows the spectrum distribution and cluster centres in a 3D space for Data_4.

The landslide at Tsao-Ling was caused by numerous blocks sliding along the bedding planes between sandstone and shale at the southwestern side of the mountain during the Chi-Chi earthquake. The collapsing rocks due to vertical joints in the rock masses of the cliff and debris rolled and slid down the slope and into the valley (Water Conservancy Agency, 1999; Yang et al., 2004). Estimation and identification of the landslide area were essential before rescue and hazard mitigation were undertaken.

Results and analysis

Figure 8 shows that the superior fitness varies with the assigned cluster numbers for Data_3 and Data_4. Four clusters for Data_3 and three clusters for Data_4 were found by the GA classifier. **Table 2** shows classification results for Data_3 through ISODATA by assigning the cluster number of four and the GA automated classifier. Very similar cluster centres were found by both classifiers. Through the GA classification, **Figure 9** shows a classified dolphin image that was identified as four clusters, including the dark black representing the tracing line of the dolphin, the light black representing the dolphin's back, the white representing the dolphin's belly, and the grey representing the background.

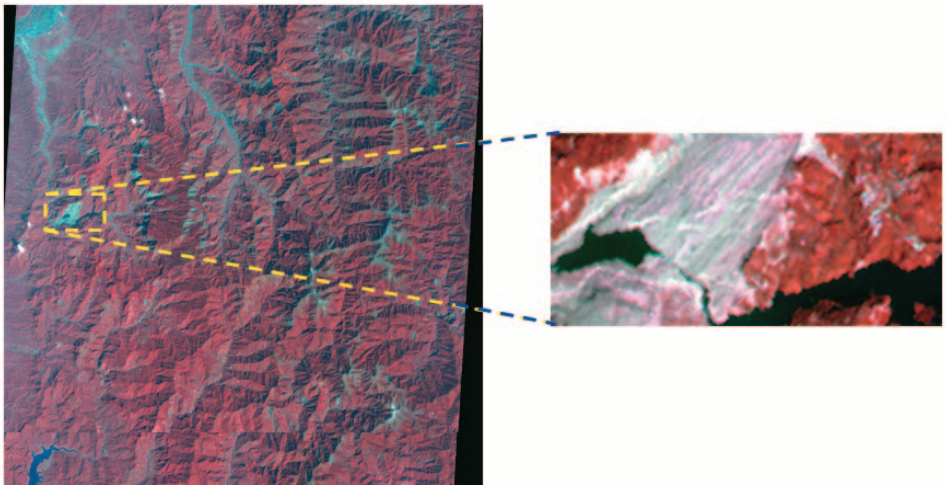

Figure 5. SPOT4 satellite image of the Tsao-Ling area (rectangular area as Data_4).

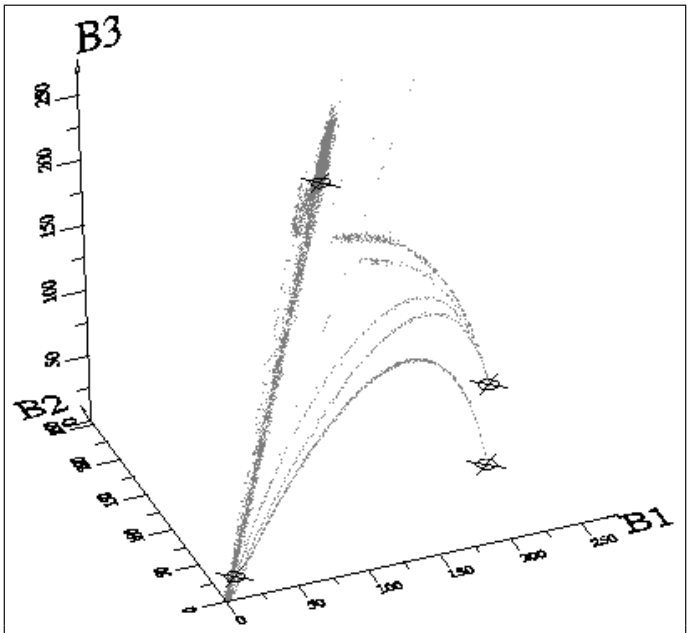

Figure 6. Distribution of pixels and cluster centres for Data_3 (with R as B1, G as B2, and B as B3).

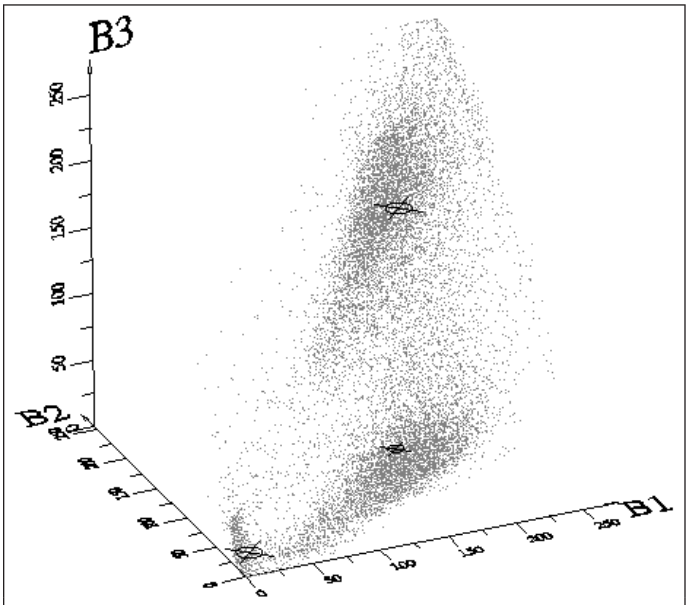

Figure 7. Distribution of pixels and cluster centres for Data_4.

Under the assignment of three cluster numbers for Data_4, ISODATA classification was run by merging the most extremely detailed 256 classes with the three most concise classes. **Table 3** is a list of the classification results for Data_4, including the computed cluster numbers and the classified cluster centres. **Figures 10** and **11** show the classifications in which three clusters were yielded for the satellite image by ISODATA and the GA classifier. To judge which classifier gave more accurate clustering, error matrixes are calculated in **Table 4**. Because of the lack of ground truth, a result from the supervised classification on the SPOT image was used as reference data by training the classifier with the spectral patterns of three classes (water, forest, and landslide) that were manually identified. In an error matrix, these diagonal entries are the number of correctly classified pixels that provide the overall accuracy. The ISODATA classification of the Tsao-Ling imagery has an overall accuracy of only 87.71%, whereas the GA classification has an overall accuracy of 96.70%. Comparing **Figure 10** with **Figure 11**, a clear difference can be seen in the lower right corner where many pixels classified as water by ISODATA were classified as forest by the GA classifier. This inconsistency can be explained by **Table 4**, in which the ISODATA classification has a lower accuracy in the landslide category because of misassignment of water to landslide, especially in the southeast area, and a significant commission error of the water category caused by misassignment of forest to water in the northeast. It was proven that the GA classifier performed almost as well as a supervised classifier, but without the need for analyst knowledge of the ground or model parameter being set a priori.

The statistical coefficient, $\hat{K}$, is an extant indicator of the extent to which the percentage of correct outcomes of an error

Table 2. Classification results for Data_3.

No. of actual clusters	No. of classified clusters	Cluster centres	
		ISODATA	GA classifier
Unknown	4	{(12.34, 11.78, 13.76) (131.61, 155.76, 205.01) (189.69, 12.62, 52.40) (244.32, 161.18, 22.26)}	{(5.95, 5.58, 6.45) (131.10, 153.39, 201.42) (190.00, 12.10, 52.08) (245.16, 161.36, 21.35)}

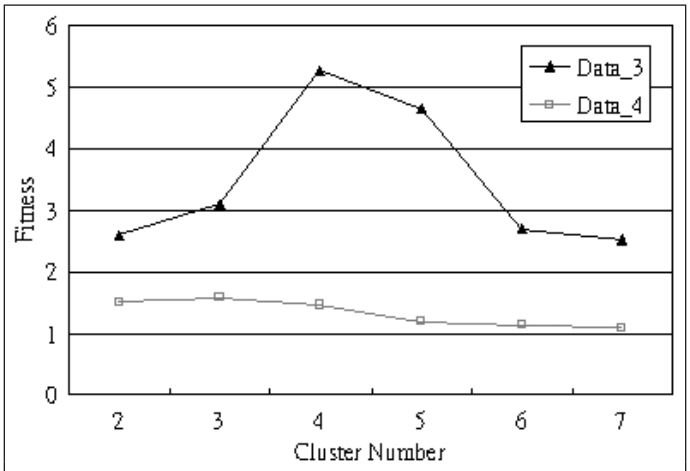

Figure 8. Superior fitness versus cluster numbers for Data_3 and Data_4.

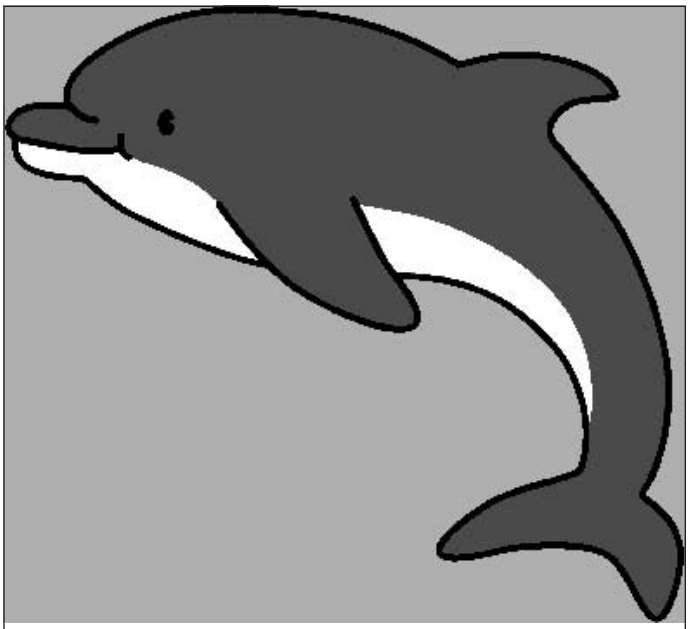

Figure 9. GA classification for Data_3.

matrix are due to "true" agreement versus "chance" agreement. In reality, the value of $\hat{K}$ is usually between 0 and 1 (Lillesand and Kiefer, 2000), where a value of 0 means that a given classification is no better than a random assignment of pixels, and a value approaching 1 means an ideal case. $\hat{K}$ can be calculated from the following equation (Lillesand and Kiefer, 2000) and is equal to 0.802 and 0.945 for the ISODATA and GA classifications for Data_4, respectively (see **Table 4**):

$$\hat{K} = \frac{N\sum_{k=1}^{K} n_{kk} - \sum_{k=1}^{K}(n_{k+} \times n_{+k})}{N^2 - \sum_{k=1}^{K}(n_{k+} \times n_{+k})} \quad (6)$$

where K is the number of classes in an error matrix, n_{kk} is the number of observations in row k and column k (the major diagonal), n_{k+} is the total number of observations in row k (shown as the row total in the matrix), n_{+k} is the total number of observations in column k (shown as the column total in the matrix), and N is the total number of observations included in the matrix.

Modification of the GA-based classifier

In the previous GA classifications, elitism was adopted by copying the best chromosome obtained in the previous iteration into the current population so that the DB index can only decrease or remain the same with an increase in generations. For insight into the convergence of genetic evolution, the GA classification was rerun by preserving the best chromosome outside the population so that the best result of each generation can be reported during the whole searching process. It was found that there was a severe disturbance during the genetic evolution because of the inconsistency between the clustering criterion and the DB index. Some previous studies reported similar results, namely that the *K*-means algorithm may fail to converge to a local minimum under certain conditions (Tseng and Yang, 2001). In the previous GA-based and ISODATA classifications, Euclidean distances between pixels and different cluster centres were calculated to determine which class is the nearest neighbor for every pixel, the so-called minimum distance (MD) classification. The GA–MD classifier has fewer provisions for accommodating differences in variability of classes for some classes that may overlap at their edges in remote sensing imagery. Thus, the clustering criterion was modified from a hard decision (minimum distance) to a soft decision by considering stochastic probability. Maximum likelihood (ML) was used in the clustering decision rule in which the nearest cluster centre is considered in relative likelihoods. Based on the assumption that all classes display multivariate normal (Gaussian) frequency distributions, the likelihood-estimate-based clustering was performed by computing the posterior probabilities of all classes.

Table 3. Classification results for Data_4.

No. of actual clusters	No. of classified clusters	Cluster centres		
		ISODATA	MD–GA classifier	ML–GA classifier
Unknown	3	{(14.64, 21.15, 25.64) (61.52, 62.96, 140.27) (171.15, 166.81, 190.35)}	{(8.55, 16.08, 7.82) (45.67, 47.87, 130.92) (165.71, 161.73, 186.17)}	{(9.19, 16.55, 10.36) ±21.69[a] (47.42, 49.44, 133.30) ±44.68[a] (167.18, 163.19, 186.67) ±58.76[a]}

[a]Standard deviation.

Table 4. Error matrix of GA classification versus ISODATA for Data_4.

	Reference data			
Classification	Water	Landslide	Forest	Row total
Water	2002 (2150)	0 (12)	0 (441)	2 002 (2 603)
Landslide	0 (8)	7198 (6066)	269 (260)	7 467 (6 334)
Forest	250 (94)	1 (1121)	6032 (5600)	6 283 (6 815)
Column total	2252	7199	6301	15 752
	Producer's accuracy (%)	User's accuracy (%)		
Water	88.90 (95.47)	100.00 (82.60)		
Landslide	99.99 (84.26)	96.40 (95.77)		
Forest	95.73 (88.87)	96.01 (82.17)		
Overall accuracy (%)	96.70 (87.71)			

Note: Values in parentheses are for the ISODATA classification.

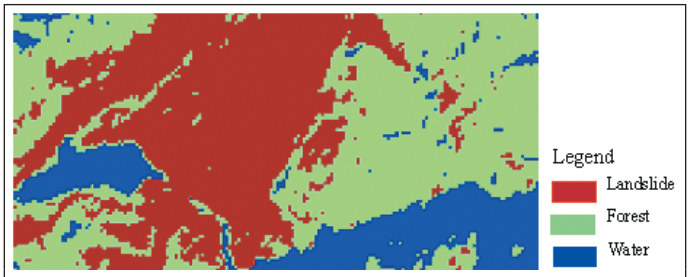

Figure 10. ISODATA classification for Data_4.

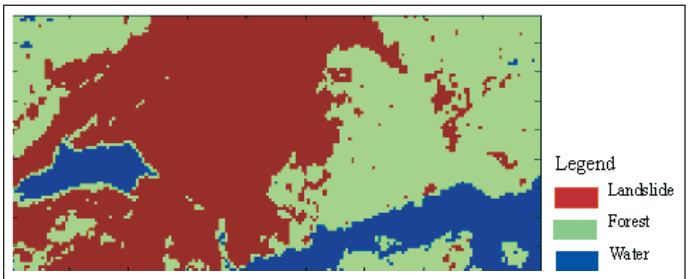

Figure 11. GA classification for Data_4.

Besides the cluster centres, the variance of clusters is essential in ML classification. The ML–GA classifier categorizes the image by minimum distance first to generate the standard deviations for all clusters in each generation as the training data, and then classifies pixel by pixel based on the probabilities (represented by a z score, a ratio of the Euclidean distance to the standard deviation) with all cluster centres. The crisp decision philosophy of winner takes all was used to choose the nearest centre of a cluster where the pixel has a maximum probability (a minimum z score) of belonging to a cluster. Thus, after the dataset is clustered by Equations (1) and (2), clustering must be run by replacing the membership function Equation (1) by Equation (7) to obtain the DB index:

$$u_{ij} = \begin{cases} 1 & \min\limits_{j=1 \text{ to } K} \dfrac{\|x_i - v_j\|}{S_j} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The ML–GA classifier was tested on Data_1 and Data_2 and obtained the cluster number and cluster centres with 100% accuracy. A better fitness value was found for the remote sensing image by the ML–GA classifier in **Figure 12**. The best DB index was almost attained by the ML–GA classifier at around the 15th generation and completely attained at the 25th generation, which was earlier than that by the MD–GA classifier. Also, the disturbance during the genetic evolution had been significantly decreased by the ML–GA classifier (see **Figure 13**). An error matrix in **Table 5** was estimated by executing a supervised classification using ML as a clustering criterion for comparison. The ML–GA classification with an overall accuracy of up to 99.51% and $\hat{K}$ of 99.20% was proven to perform as well as a supervised classifier. Based on the quantitative analyses, the ML–GA has the following advantages: rapid approach to optimal clustering, suitability for remote sensing images with class overlap, satisfactory accuracy, and total automation.

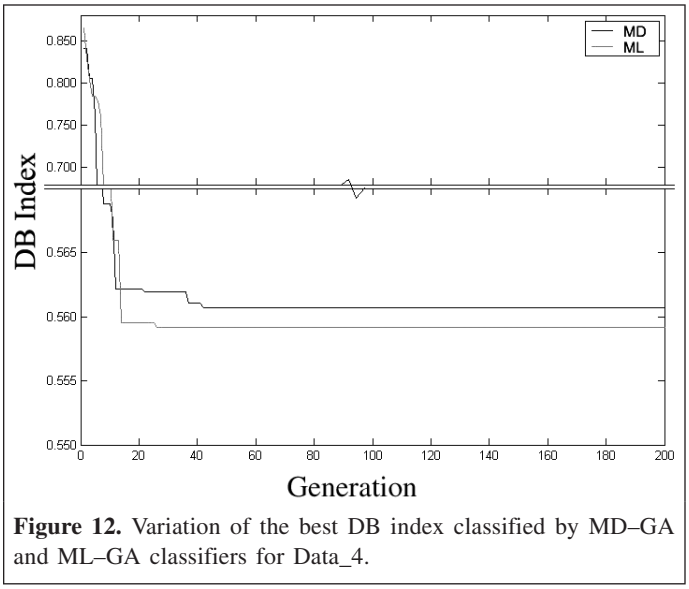

Figure 12. Variation of the best DB index classified by MD–GA and ML–GA classifiers for Data_4.

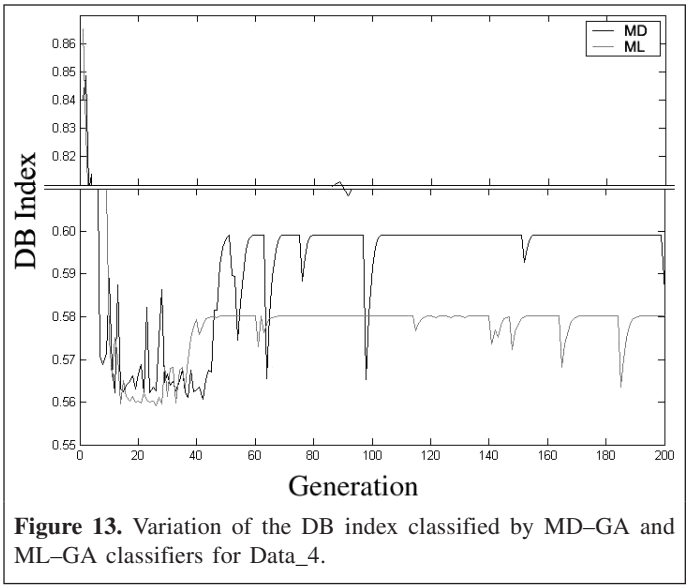

Figure 13. Variation of the DB index classified by MD–GA and ML–GA classifiers for Data_4.

Table 5. Error matrix of ML-GA classification for Data_4.

	Reference data			
Classification	Water	Landslide	Forest	Row total
Water	2199	0	0	2 199
Landslide	0	7191	16	7 207
Forest	53	8	6285	6 346
Column total	2252	7199	6301	15 752
	Producer's accuracy (%)	User's accuracy (%)		
Water	97.65	100.00		
Landslide	99.89	99.78		
Forest	99.75	99.04		
Overall accuracy (%)	99.51			

Conclusions

In classical unsupervised clustering methods, determination of the cluster number and improvement of clustering accuracy need more effort, especially where there is little knowledge of ground truth. This paper proposes an automated unsupervised classifier using a genetic algorithm (GA) that was tested on two artificial datasets and applied to two real images. The effectiveness of the GA classifier was assessed by the artificial data, and the results verified that the GA classifier is able to classify these datasets into exact clusters and accurately locate cluster centres for artificial data without any information a priori. For application to real digital imagery, the GA classifier performed more satisfactorily than the ISODATA on both the dolphin picture and the satellite image of the Tsao-Ling landslide area. Without ground truth, the performance of ISODATA and GA classifications was assessed by comparing the result with a supervised classification on the SPOT4 satellite image. The overall accuracies have validated that the GA classifier (96.70%) was superior to ISODATA (87.71%). Also, the GA classification with a higher $\hat{K}$ (0.945) than that (0.802) of ISODATA means the GA classifier is more ideal than ISODATA in the classification of the Tsao-Ling satellite image. Furthermore, a modified GA classifier with a z score as a clustering criterion performed more accurately and more robustly in the application of satellite data. This modification not only decreases the disturbance of the GA searching process because of the inconsistency between the clustering criterion and clustering validity, but also approaches a better clustering result with a higher overall accuracy (99.51%). It has been proven that the GA-based automated classifier is able to classify remote sensing imagery, which seldom records spectrally pure classes and often has an overlap of classes, with a high accuracy and without human effort. Future work includes enhancing the GA-based classifier with more versatile and feasible applications for automatic classification on remote sensing imagery.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan, Republic of China, under grant NSC-92-

2211-E-005-036. My graduate students, Y.Y Yang and C.H. Hsu, are appreciated for their help in program coding. Thanks also go to the reviewers for their constructive suggestions.

References

Alippi, C., and Cucchiara, R. 1992. Cluster partitioning in image analysis classification: a genetic algorithm approach. In *CompEuro '92, Proceedings of the IEEE International Conference on Computer Systems and Software Engineering*, 4–8 May 1992, The Hague. IEEE Computer Society Press, Los Alamitos, Calif. Vol. 4, pp. 139–144.

Avery, T.E., and Berlin, G.L. 1992. *Fundamentals of remote sensing and airphoto interpretation*. 5th ed. MacMillan Publishing Company, New York.

Bandyopadhyay, S., and Maulik, U. 2002. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition*, Vol. 35, No. 6, pp. 1197–1208.

Bezdak, J.C., and Pal, N.R. 1998. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics*, Vol. 28, pp. 301–315.

Bischof, H., Leonardis, A., and Selb, A. 1999. MDL principle for robust vector quantization. *Pattern Analysis & Applications*, Vol. 2, pp. 59–72.

Chipperfield, A., Fleming, P., Pohlheim, H., and Fonseca, C. 1994. *Genetic algorithm toolbox version 1.2: user guide for use with MATLAB*. Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, UK.

Dave, R.N., and Krishnapuram, R. 1997. Robust clustering methods: a unified view. *IEEE Transactions on Fuzzy Systems*, Vol. 5, pp. 270–293.

Diederich, J., and Fortuner, R. 1999. A fuzzy classifier using genetic algorithms for biological data. In *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, 10–12 June 1999, New York. IEEE, New York. pp. 680–684.

Garai, G., and Chaudhuri, B.B. 2004. A novel genetic algorithm for automatic clustering. *Pattern Recognition Letters*, Vol. 25, pp. 173–187.

Gen, M., and Cheng, R. 1997. *Genetic algorithms and engineering design*. John Wiley & Sons, New York.

Groenen, P.J.F., and Jajuga, K. 2001. Fuzzy clustering with squared Minkowski distances. *Fuzzy Sets and Systems*, Vol. 120, pp. 227–237.

Guliato, D., Rangayyan, R.M., Camielli, W.A., Zuffo, J.A., and Leo Desautels, J.E. 2003a. Segmentation of breast tumors in mammograms using fuzzy sets. *Journal of Electronic Imaging*, Vol. 12, pp. 369–378.

Guliato, D., Rangayyan, R.M., Camielli, W.A., Zuffo, J.A., and Leo Desautels, J.E. 2003b. Fuzzy fusion operators to combine results of complementary medical image segmentation techniques. *Journal of Electronic Imaging*, Vol. 12, pp. 379–389.

Lillesand, T.M., and Kiefer, R.W. 2000. *Remote sensing and image interpretation*. John Wiley & Sons, New York.

Martini, H., and Schobel, A. 2001. Median and center hyperplanes in Minkowski spaces — a unified approach. *Discrete Mathematics*, Vol. 241, pp. 407–426.

Maulik, U., and Bandyopadhyay, S. 2003. Fuzzy partition using a real-coded variable-length genetic algorithm for pixel classification. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 41, No. 5, pp. 1075–1081.

Miller, D.M., Kaminsky, E.J., and Rana, S. 1995. Neural network classification of remote sensing data. *Computers & Geosciences*, Vol. 21, pp. 377–386.

Mitra, S. 2004. An evolutionary rough partitive clustering. *Pattern Recognition Letters*, Vol. 25, pp. 1439–1449.

Mohanty, K.K., and Majumdar, T.J. 1996. An artificial neural network (ANN) based software package for classification of remotely sensed data. *Computers & Geosciences*, Vol. 22, pp. 81–87.

Murthy, C.A., and Chowdhury, N. 1996. In search of optimal clusters using genetic algorithm. *Pattern Recognition Letters*, Vol. 17, pp. 825–832.

Murthy, C.S., Raju, P.V., and Badrinath, K.V.S. 2003. Classification of wheat crop with multi-temporal images: performance of maximum likelihood and artificial neural networks. *International Journal of Remote Sensing*, Vol. 24, pp. 4871–4890.

Nascimento, S., Mirkin, B., and Moura-Pires, F. 2003. Modeling proportional membership in fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, Vol. 11, pp. 173–186.

Oin, A.K., and Suganthan, P.N. 2004. Robust growing neural gas algorithm with application in cluster analysis. *Neural Networks*, Vol. 17, pp. 1135–1148.

Ostir, K., Veljanovski, T., Podobnikar, T., and Stancic, Z. 2003. Application of satellite remote sensing in natural hazard management: the Mount Mangart landslide case study. *International Journal of Remote Sensing*, Vol. 24, pp. 3983–4002.

Pakin, S.K., Gaborski, R.S., Barski, L.L., and Parker, K.J. 2003. A clustering approach to bone and soft tissue segmentation of digital radiographic images of extremities. *Journal of Electronic Imaging*, Vol. 12, pp. 40–49.

Pierce, L., Samples, G., Dobson, M.C., and Ulaby, F. 1998. An automated unsupervised/supervised classification methodology. In *IGARSS'98, Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, 6–10 July 1998, Seattle, Wash. IEEE, New York. Vol. 4, pp. 1781–1783.

Ross, T.J. 1995. *Fuzzy logic with engineering applications*. McGraw-Hill, New York.

Rydberg, A., and Borgefors, G. 2001. Integrated method for boundary delineation of agricultural fields in multispectral satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 39, pp. 2514–2520.

Swanepoel, K.J. 1999. Cardinalities of k-distance sets in Minkowski spaces. *Discrete Mathematics*, Vol. 197, pp. 759–767.

The Math-Works, Inc. 2002. *MATLAB® — the language of technical computing, version 6.5*. The MathWorks, Inc, Natick, Mass.

Tsai, D.Y., Lee, Y., Sekyia, M., and Ohkubo, M. 2004. Medical image classification using genetic algorithm based fuzzy-logic approach. *Journal of Electronic Imaging*, Vol. 13, pp. 780–788.

Tseng, L.Y., and Yang, S.B. 2001. A genetic approach to the automatic clustering problem. *Pattern Recognition*, Vol. 34, No. 2, pp. 415–424.

van der Sande, C.J., De Jong, S.M., and de Roo, A.P.J. 2003. A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk and flood damage assessment. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 4, pp. 217–229.

Water Conservancy Agency. 1999. *The final report of treatment on 921 earthquake area — Tsao-Ling area*. Water Conservancy Agency, Ministry of Economic Affairs, Taipei. [In Chinese.]

Xie, X.L., and Beni, G. 1991. A new fuzzy clustering validity criterion and its application to color image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, pp. 841–847.

Yang, M.D., and Su, T.C. 2006. An automation model of sewerage rehabilitation planning. *Water Science and Technology*, Vol. 54, No. 11–12, pp. 225–232.

Yang, M.-S., and Wu, K.-L. 2001. A new validity index for fuzzy clustering. In *Proceedings of the 10th IEEE International Fuzzy Systems Conference*, 2–5 December 2001, Melbourne, Australia. IEEE, New York. pp. 89–92.

Yang, M.D., and Yang, Y.F. 2004. Genetic algorithm for unsupervised classification of remote sensing imagery. In *Image Processing: Algorithms and Systems III*. Edited by E.R. Doughherty, J.T. Astola, and K.O. Egiazarian. Proceedings of SPIE Vol. 5298, pp. 395–402.

Yang, M.D., Yang, Y.F., and Hsu, S.C. 2004. Application of remotely sensed data to the assessment of terrain factors affecting the Tsao-Ling landside. *Canadian Journal of Remote Sensing*, Vol. 30, No. 4, pp. 593–603.

Yang, M.D., Su, T.C., Hsu, C.H., Chang, K.C., and Wu, A.M. 2007. Mapping of the 26 December 2004 tsunami disaster by using FORMOSAT-2 images. *International Journal of Remote Sensing*, Vol. 28, Nos. 13–14, pp. 3071–3091.

Zhang, C., and Wang, F. 1994. A genetic algorithm for training image classification neural networks. In Humans, Information, and Technology: 1994 IEEE International Conference on Systems, Man, and Cybernetics, 2–5 October 1994, San Antonio, Tex. IEEE, New York. Vol. 3, pp. 2242–2247.