GENETIC ALGORITHMS OPTIMIZATION FOR LAND-COVER CLASSIFICATION FROM HIGH RESOLUTION DIGITAL IMGERY

Mahmoud Al-Nokrashy Osman(1), Adel Ahmed Esmat(1),Mahmoud Salah Mahmoud(2), Ahmed Mohamed Hamdy(3)

(1) Professor of Surveying, Dept. of Civil Engineering, Al-Azhar University.
(2) Lecturer of Surveying, Dept. of Surveying, Shoubra, Benha University,
(3) Assistant Lecturer, Dept. of Civil Engineering, Al-Azhar University.

ملخص البحث:

يزداد الاهتمام بدراسة المناطق الحضرية باستخدام الصور الجوية الرقمية عالية الدقة. ولقد زادت الصور الرقمية عالية الدقة من صعوبة عملية التصنيف الاوتوماتيكي نتيجة تشابه المعلومات المكانية والطيفية لنفس المعلم. ولذلك تم التوصل الى تقنيات حديثة للتصنيف وانتاج الخرائط للمناطق الحضرية. هذه الدراسة تطرح طريقة جديدة للاستنتاج الاوتوماتيكي للمعالم باستخدام الخوارزمية الجينينة. وكان الهدف من البحث هو الوصول لأفضل قيم لمدخلات الخوارزمية الجينية باستخدام K-means كمصنف. وكانت أفضل نتائج التصنيف فى حالة استخدام القيم الآتية: حجم المجتمع = 100 ونسبة احتمالية الطفرة = 0.05 بدقة كلية مقدارها 68.89%.

KEY WORDS: Digital Imagery, Unsupervised Classification, Genetic Algorithm, K-means Index.

ABSTRACT:

Study of urban environmental areas involved with the use of digital imagery data has raised great interest among researchers. High resolution imagery present difficulties for automatic classification process due to the high spectral and spatial heterogeneity for the same class. Thus, new concepts and techniques have been used for mapping urban areas. In this study Genetic Algorithms (GAs) were applied to determine the optimal input parameters based on k-means classifier as a fitness function. To assess the efficacy of the methodology and ensure the accuracy of the product the steps undertaken in this study were subject to quality control. The best results were obtained in the case of Population size 100 with mutation probability 0.05 with overall accuracy of 68.89%.

1. INTRODUCTION

Land-cover classification including supervised and unsupervised classification is one of the important applications of digital imagery. The main goal of the image classification is the associations of each pixel in the image with a specific land-cover class to produce accurate classification maps from the data. Unsupervised classification divides pixels within an image into a corresponding cluster pixel by pixel. Typically, the only input is the number of clusters of the scene. In general, unsupervised classification is performed through the well-known method, k-means (waske, 2007). Heuristic unsupervised classification works by establishing some mathematical models and then determining the cluster numbers and centroids automatically by optimizing the model parameters to obtain higher accuracy (Yang, 2006).

On the other hand, supervised classification can be performed through a variety of methods such as classification trees, neural network and machine learning. The major steps of supervised classification may include: determination of suitable classification system; selection of training samples; image preprocessing; feature clustering; post-classification processing; and accuracy assessment (Lu and Weng, 2007).

Genetic algorithms (GAs), introduced by John Holland in 1975, are suitable method to produce heuristic unsupervised classification (Coley, 1999; Pham and Karraboga, 2000). Image classification has been widely and successfully applied by optimization algorithms specifically GAs(Rothlauf, 2006), which confirm the potential of GAs to produce high level of quality results especially when applied without any ground truth (Coley, 1999).Numerous studies have shown that the GAs technique is very efficient in dealing with large datasets and has a large chance to avoid a local optimal solution than other methods (Huang et al. 2006, Zhou et al. 2010). Another advantage of the GAs is its capability to search for input features and parameters of classifier simultaneously.

Almeida(2012)studied the use of GAs routine with decision trees for the object-based land-cover classification. The study showed a satisfactory performance for the automatic assessment of the optimal segmentation parameters. Nevertheless, the shape complexity of some targets, the internal spectral variability of certain classes, and the diverse conditions of ageing and maintenance of some roof classes found in the study area led to an over-segmentation of some targets. Adding height information derived from laser scanning to the imagery discriminates targets with similar spectral behavior but diverging values.Chu (2012) used the integration of feature selection using GAs and multi classifier system with Dempster-Shafertheory for classifying different combined datasets. Results of classification revealed that proposed method (FS-GA-DS model) always gave significantly higher accuracy, than any single classifier.Ge (2012) proposed GA-SVM model to classify multiple combined datasets, consisting of Landsat 5 TM, Multi-date dual polarization ALOS/POLSAR images and their multi-scale textural information. The performance of the proposed method was compared with that of the traditional steck-vector approach. It revealed that the proposed method is efficient for handling multisource data. The highest classification accuracy achieved was 96.47% with only 81 out of 189 features being selected which demonstrate the advantages of using multi-source data over single source data.Li (2013) introduced GAs and SVM to high resolution POLSAR image classification. This model was applied with the contrast of three classifiers without using additional polarimetric information, and thus its overall accuracy is only 74.85%. The method gets the best results when using additional information, the accuracy is up to 97.49%.Yang (2006) studied the influence of changing GAs parameters on classification results. Two of these parameters, namely population size and the crossover probability were considered. He pointed out in his results that the population size proofed to be significantly more important than the crossover probability. The effectiveness of this technique was evaluated using IKONOS satellite images. An overall accuracy of 71.1% was reached using (DBI) index as compared to 65.1% when using the ISODATA algorithm.Jamshidpour (2012) suggested a framework to combine filter and wrapper feature selection methods to find feature subset and optimize the SVM kernel parameters at the same time.GAs have been used as global optimizer to obtain the optimal solution after a series of iterative computations.Samaher(2007) attempted to classify 6 different kinds of forest scenes using genetic algorithms clustering and neural network. The randomly estimation of the number of the clusters that are found in the image may lead to error in classification process. Therefore, the proposed methodology can solve this problem by determining it automatically. Optimal results will depend on the selection of parameters.

The main objective of this research is the heuristic optimization process of GAs to produce an accurate, time-effective and automatic method to classify very high resolution digital imagery. In this regard, GAs with K-means index is adopted to study the influence of changing GAs parameters on determining the number of clustersas well astheaccuracy of the classificationresults.All the methods proposed in this research were implemented inMatlab environment.

2. PRINCIPLES OF GENETIC ALGORITHM

Genetic Algorithms are adaptive methods which may be used to solve a variety of optimization problems by the principles of biological organisms' evolution. Over the last twenty years, it has been used to solve a wide range of search and optimization problems specifically in unsupervised classification of digital imagery (Bandyopadhyay and Maulik, 2002; Majida, 2010).GAs is very different from most of the traditional optimization methods. GAs generates a population of solutions at each iteration to approach an optimal solution.This means that GAs can processa number of designs at the same time. In addition it selects the next population by computations that involve random choices. Figure 1 summarizes the working principles of GAs. The following sections describe the general operations of GAs.

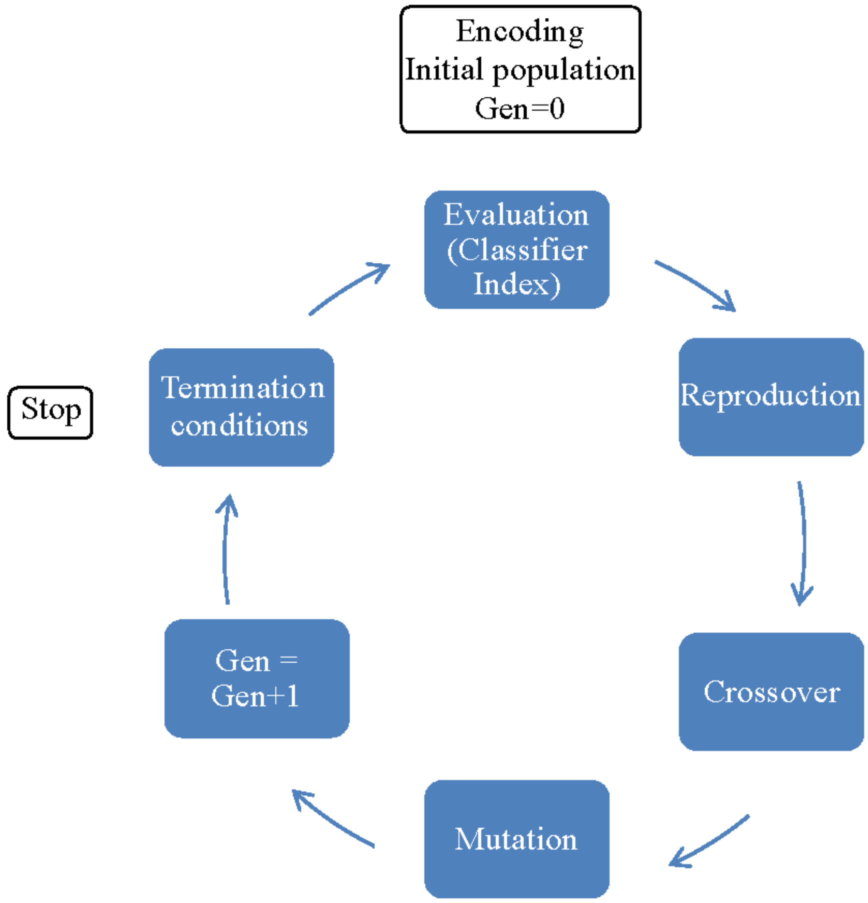

Figure (1): Major steps of image classification using Genetic Algorithm.

2.1. Initialization

There should be an initial population of individuals to perform genetic operations. It is chosen according to the problem domain and solution strategy; the population can be generated randomly or obtained from a training data. Commonly, it is generated randomly, covering all possible solutions. Occasionally, the solutions may be coded in the area where optimal solutions are likely to be found. The typical size of the population can range from 20 to 1000 (Coley, 1999). Thefollowing is example to explain the creation of aninitial population: Digital image withthree bands was assumed. $K_{min}$is set to 2 and $K_{max}$to 8, where K is the length of the chromosome. In Genetic Algorithms, the unknown parameters are encoded in the form of strings, so-called chromosomes.At the beginning, foreach chromosome i (i =1, 2,…,.P, where P is the size ofpopulation) all values are chosen randomly from the data space. Such a chromosomebelongs to the so-called parent generation. For example one chromosome of the parent generation is given here:

-1 (101, 56, 234) (170, 56, 216) -1 (22, 119, 4) (110, 50, 210) -1 (217, 150, 92)

2.2. Reproduction

The genetic encoding representation can be decided according to the problem domain. There are several types of encoding such as; binary, real and integer encoding. Since multispectral image data are usually represented by positive integers, in this research a chromosome is encoded with a unit of positive integer numbers. Each unit represents a combination of brightness values, one for each band, and thus a potential cluster centroid. K, the length of the chromosome, is equivalent to the number of clusters in the classification problem. K is selected from the range (K_{min}, K_{max}), where K_{min} is usually assigned to 2 unless special cases are considered (Bandyopadhyay and Maulik, 2002), and K_{max} describes the maximum chromosome length, which represents the maximum number of possible cluster centroids. K_{max} must be selected according to experience. Invalid (non-existing) clusters are represented with negative integer "-1". The values of the chromosomes are changed in an iterative process to determine the correct number of clusters (the number of valid units in the chromosomes) and the actual cluster centroids for a given classification problem.

Reproduction (or selection) is usually the first operator applied to a population. The commonly used reproduction operator is the proportionate selection operator, where a string in the current population is selected with probability proportional to the string's fitness. Thus, the i^{th}string in the population is selected with probability proportional tof_i . Since the population size is usually kept fixed in a simple GAs, the cumulative probability for all string in the population must be one. Therefore, the probability for selecting i^{th}string is i $f_i = \sum_{i=1}^{N} f_i$, where N is the population size. One way to achieve this proportionate selection is to use a roulette-wheel with the circumference marked for each string proportionate to the string's fitness.After selecting fittest parents for reproduction processes; crossover, and/or mutation, new child individual(s) are reproduced from each selected pairs of parents. The process that enables gene exchange between parents is defined as "crossover" which creates two new different individuals from the existing parents. Also, mutation is the process that provides a random/rule based gene change on the individuals.

2.2.1. Crossover

Crossover operator specifies how the genetic algorithm combines two individuals, to form a crossover child for the next generation.The crossover mission is to create two new individual chromosomes from two existingchromosomes selected randomly from the current population.Typical crossover operations are one-point crossover, two-point crossover, cycle crossover and uniform crossover. In thisresearch, the two-point crossover was adopted due to the length of the chromosome.

2.2.2. Mutation

After the crossover is carried out, mutation takes place. The algorithm creates mutation children by randomly changing the values of some genes location in the chromosome. This operator may be implemented by simple way such as flip bit or by other ways such as boundary, non-uniform, uniform and Gaussian.

2.3. Indices Identification

The genetic algorithm should succeed two goals: maximizing the classification accuracy, and minimizing the number of selected features. These criteria used to create a single objective function as follows:

$$F = w * C(x) + (1\text{-}w) * \frac{1}{N(x)} \tag{1}$$

Where x is the feature subset, $C(x)$ represents the classification accuracy, $N(x)$ is the size of selected feature subset, and w is a parameter between 0 and 1 which adjusts the influence of each criterion. As value of w is higher the weight of classification accuracy in fitness function is greater. On the other hand, reducing the value of w will give more penalties on the size of$x$(Tan, Fu et al. 2008). By adjusting w, we can achieve a trade-off between the accuracy and the size of the feature subset obtained. For this research, w was adjusted to 0.8 to avoiding large decrease in classification accuracy.

Based on crossover and mutation operations, the process, iteratively evolve from one generation to the next. Inorder to be able to stop this iterative process, a so-called fitnessfunction needs to be defined to measure the fitness oradaptability of each chromosome in the population. Previous research used different indices, such as distance, separation index, Fuzzy C-Means Index, K-means Index (KMI), Davies-BouldinIndex (DBI), and Xie-Beni Index (XBI), as criteria to determinethe best clustering. For this research, the K-means was adopted, because it is not as complex as other classifiers and one can obtain better results than with some other indices (Yang and Wu, 2001; Bandyopadhyay and Maulik, 2002).

2.4. Termination

Termination conditions control the process of reproducing new generations. The genetic algorithm uses some common conditions to stop the process such as: reaching the value of fixed number; if there is no significant improvement in the objective function; after running for an amount of time in seconds equal to time limit and other conditions.

3. METHODOLOGY

3.1. GA Application for Unsupervised Classification

There are basically seven parameters that influence the result of classification using Genetic Algorithm. These parameters include: the maximum length of the chromosome, the way to encode the chromosome units (binary, real number and so on), the population size, the crossover type and probability, the mutation probability, and the employed fitness function (Pham and Karaboga, 2000).

3.2. Test Description

In this research, a maximum chromosome length of K_{max}=8 was chosen, which is above the maximum number of clusters in the test image. As mentioned above, chromosome coding was done using positive integers. Only two-point crossover operations were considered with fixed probability. The other parameters were systematically varied in order to study their influence on the result. For selecting the actual parameter values we took advice from general GA references (Coley, 1999). More specifically, the population size was set to 20, 30, 40:200 chromosomes, respectively. Mutation probability values were set to 0.005, 0.05, and 0.5 then to 0.001, 0.003: 0.009 then to 0.01, 0.03:0.09. One set of parameters, namely a population size of 100with a mutation probability of 0.005 was considered as the baseline set, against which the other parameters were varied. Iterations were terminated as soon as they arrive the 100 generation. In this way, the K-means index as a fitness functionwasused and investigated, resulting in a wide variety of results.

4. THE FITNESS FUNCTION (KMI)

For the iteration process, the fitness function (index) is used to measure the fitness or adaptability of each chromosome in the population. The best obtained chromosome is compared to the best one of the previous iteration after calculating the index for each chromosome of a given population.K-Means is a simple and common clustering algorithm which can also be used within GAs framework. KMI represents the total variationdisregarding the distance between different clusters. KMI is computed as follows:

$$KMI = 1/(\sum_{K=1}^{K} \quad \sum_{i=1}^{N} \mu_{ik} \, ||x_i - v_k||^2 \tag{2}$$

K = total number of clusters

N = total number of pixels

μ_{ik}= membership function of each pixel x_i belonging to the K^{th}cluster

x_i= pixel i with grey values x (one for each band)

v_k= average value of K^{th}cluster in the current iteration

5. STUDY AREA AND DATA PROPERTIES

For this research a very high resolution digital image was available.The imagecovers an area of approximately 500x500m of the region surrounding the University of New South Wales (UNSW)campus, Sydney Australia.The area is a largely urban area that contains residential buildings, large Campus buildings, and a network of main roads as well as minor roads, trees, open areas and green areas.The color imagery was captured by film camera at a scale of 1:6000. The film was scanned in three color bands (red, green and blue) in TIFF format, with 15µm pixel size (GSD of 0.09m) and radiometric resolution of 16-bit as shown in Figure 2. The characteristics of image datasets are provided in Table 1.

Table (1): Characteristics of image datasets.

Test area	*Size(Km)*	*bands*	*pixel size (cm)*	*Camera*	*Look Angle (deg.)*	
					along track	across track
UNSW	0.5 x 0.5	RGB	9	LMK1000	±30	±30

In order to perform the classification process, three main object classes: trees, roads, and buildings were selected.

Figure (2): An Orthophotoof UNSW campus.

6. RESULTS AND DISCUSSION

First, the effect of population size variation on the obtained classification accuracies was tested. For each case, the Producer's Accuracy (PA), the User's Accuracy (UA), the Overall Accuracy (OA) as well as the K-HAT value are given with respect to a variety of population size (20, 50, 60…200). Table 2 displays the visual and numerical results. The colors indicate the different classes: Gray stands for *roads*, and Green for *trees, and Maroon for buildings.*Table 3 contains similar results for the variation of mutation probability (0.005, 0.05 and 0.5), (0.001, 0.003…0.009) then (0.01, 0.03….0.09).

<u>From on the obtained results the following conclusions can be drawn:</u>

- Few results showed only one class (building) in cases of Population size 130 and mutation probability 0.003 and 0.07.
- In cases of Population size 90, 120, 150, and 180,and mutation probability of 0.007, the results showed two clusters (building and roads). Population size 200 and mutation probability 0.009 and 0.01 showed similar results.
- The results were improved to show three clusters (roads, building and trees) in case of Population size 100 and mutation probability 0.05, 0.03 and 0.09.
- The best results were shown in the case of Population size 100 with mutation probability 0.05 with overall accuracy of 68.89%.
- For all cases, the K-HAT value is rather low. This indicates a number of errors of omission and commission, which can also be observed when investigating the full error matrices.These matrices will be given in Hamdy(2015).

Table (2): Numerical results for population size variation.

Probability of Mutation = 0.005 & No. of Generations = 100								
Population Size	Overall Accuracy %	Class Accuracy						K-HAT %
		Roads		Trees		Buildings		
		PA %	UA %	PA %	UA %	PA %	UA %	
20	non	non	non	non	non	non	non	non
50	non	non	non	non	non	non	non	non
60	non	non	non	non	non	non	non	non
70	non	non	non	non	non	non	non	non
80	non	non	non	non	non	non	non	non
90	33.33	100	48.39	non	non	00.00	00.00	0.1089
100	44.44	46.67	37.84	33.33	90.91	53.33	100	0.2718
110	non	non	non	non	non	non	non	non
120	16.67	00.00	00.00	non	non	50.00	27.27	-0.1250
130	12.22	non	non	non	non	36.67	15.49	-0.1910
140	non	non	non	non	non	non	non	non
150	33.33	100	33.3	non	non	non	non	00.00
180	22.22	3.33	5.56	non	non	63.33	95.00	0.0948
200	18.89	non	non	56.67	94.44	00.00	00.00	0.1275

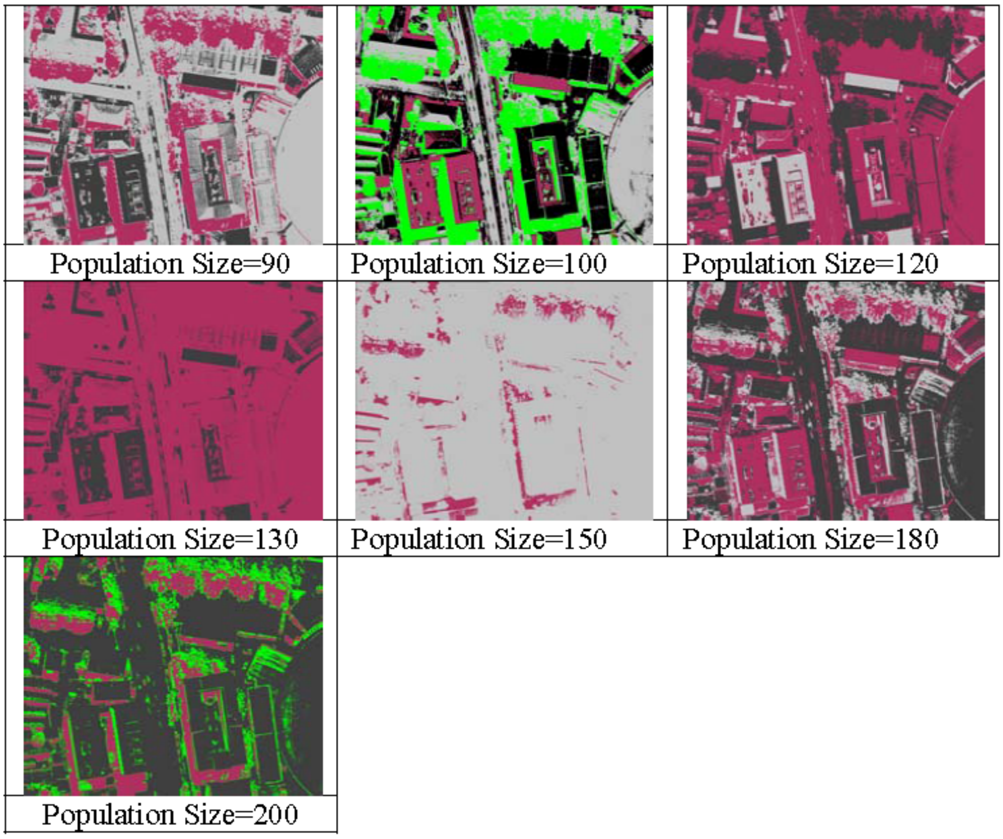

Figure (3): Typical examples of the results showing the effect of population size variation (Probability of Mutation = 0.005 & No. of Generations = 100).

Table (3): Numerical results for mutation probability variation.

Population Size = 100 & No. of Generations = 100								
Mutation Probability	Overall Accuracy %	Class Accuracy						K-HAT %
		Roads		Trees		Buildings		
		PA %	UA %	PA %	UA %	PA %	UA %	
0.005	44.44	46.67	37.84	33.33	90.91	53.33	100	0.2718
0.05	68.89	96.67	52.73	43.33	92.86	66.67	100	0.5359
0.5	00.00	non	non	00	00	00	00	-0.1638
0.001	non	non	non	non	non	non	non	non
0.003	21.11	non	non	non	non	63.33	100	0.1514
0.007	6.67	00	00	non	non	20.00	9.84	-0.3622
0.009	34.44	non	non	40.00	92.31	63.33	100	0.2563
0.01	47.87	non	non	96.67	90.63	46.67	87.50	0.3649
0.03	38.89	non	non	40.00	85.71	76.67	82.14	0.2763
0.07	00.00	non	non	non	non	00.00	00.00	-0.0345
0.09	20.00	6.67	14.29	53.33	34.78	00.00	00.00	-0.0385

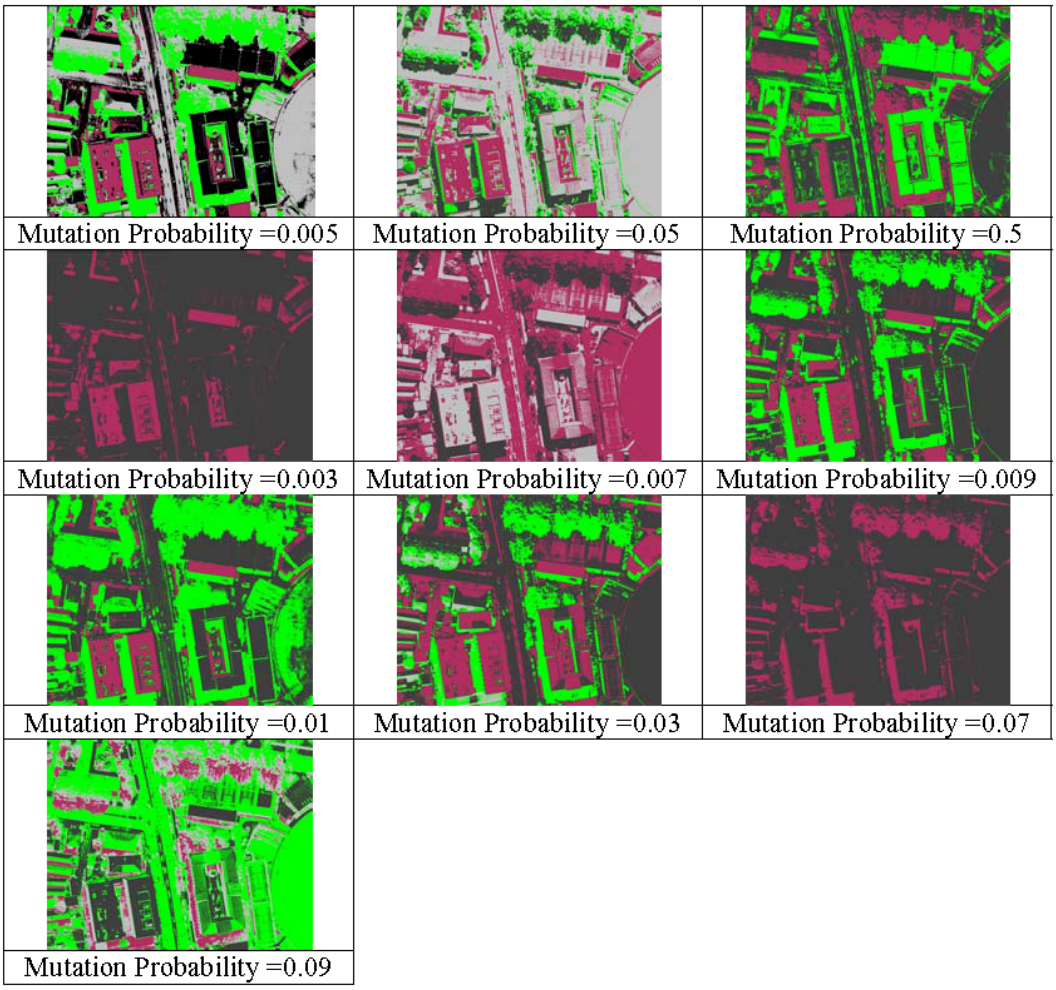

Figure (4): Typical examples of the results showing the effect of mutation probability variation (Population Size = 100 & No. of Generations = 100).

7. CONCLUSIONS

This study tried to optimize and validate GAs for unsupervised classification of very high resolution image. GA provides a possibility to compute the number of clusters present in a scene from the image data by using a particular fitness function.Experimental results were obtained by classifying a high resolution scene for a part of the region surrounding the University of New South Wales campus, Sydney Australia, depicting three different classes, namely buildings, roads and trees with the KMI as a fitness function while varying a number of parameters of the GA. Compared with the reference data, the results were evaluatedbased on a variety of criteria which includes visual inspection, error matrices and the K-HAT statistics. The results showed that, the GA-KMI model is much more sensitive for parameter tuning. The results were improved considerably in cases of Population size 100 and mutation probability 0.05, 0.03 and 0.09. The best results were obtained in the case of Population size 100 with mutation probability 0.05 with overall accuracy of 68.89%. Thus, GA algorithms seem to be more flexible and therefore advantageous to more traditional unsupervised classification techniques.In order to improve the results achieved from this study, different data sources have been extended simultaneously. Also, experiment has been done with different fitness functionsto get best results. Finally, we want to integrate indices based on fuzzy theory into the investigations. The results ofall these researches are given in Hamdy (2015).

8. ACKNOWLEDGEMENTS

The authors would like to thank the AAMHatch (www.aamhatch.com.au) and the School of Surveying and Spatial Information Systems, The University of New South Wales, Sydney, Australia for providing the multispectral aerial imagery.

REFERENCES

Almeida, C. M., 2012.*Genetic Algorithms and Data Mining Applied to Optical Orbital and Lidar Data for Object-Based Classification of Urban Land Cover,* Proceedings of the 4th GEOBIA, May 7-9, 2012 - Rio de Janeiro - Brazil. p.649.

Bandyopadhyay, S., and Maulik, U., 2002.*Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification,* IEEE pattern recognition, Vol.35, pp.1197-1208.

Chu, H. T., 2012. *Combination of Genetic Algorothm and Dempster-Shafer Theory of Evidence for Land Cover Classification Using Integration of SAR and Optical Satellite Imagery,* International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B7, 2012 XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia.

Coley, A D., 1999.*An Introduction to Genetic Algorithms for Scientists and Engineers,* World Scientific, Singapore, 188p.

Ge, L., 2012. *Application of Genetic Algorithm and Support Vector Machine in Classification of Multisource Remote Sensing Data,* International Journal of Remote Sensing Applications Sept. 2012, Vol. 2 Iss. 3, PP. 1-11.

Hamdy A., 2015. *GA for Feature Extraction from Fused Aerial Images and Other Data Sources,* Ph.D. Thesis to be submitted to Civil Eng. Dept. Azhar University, Cairo, Egypt.

Huang, C.L. and Wang, C.J., 2006. *A GA-based Feature Selection and Parameters Optimization for Support Vector Machines,* Expert Systems with Applications, 31, 231-240.

Jamshidpour, N. 2012.*Improvement of Hyperspectral Image Classification Using Genetic Algorithm for Feature Selection and SVMs Parameters Optimization,* Geomatics Department, College of Engineering, University of Tehran North Kargar St., Tehran, Iran.

Li, P. X., 2013. *High Resolution POLSAR Image Classification Based on Genetic Algorithm and Support Vector Machine,* International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-7/W1, 3rd ISPRS IWIDF 2013, 20 – 22 August 2013, Antu, Jilin Province, PR China

Lu, D. and Weng, Q., 2007. *A Survey of Image Classification Methods and Techniques for Improving Classification Performance,* International Journal of Remote Sensing, 28(5): pp. 823 - 870.

Majida A. A., Ahmad N. I. and ZubadiMatizHazi, 2010.*Pattern Recognition Using Genetic Algorithm,* International Journal of Computer and Electrical Engineering.Department of Computer science, university of Tikrit, Iraq, Vol. 2, No. 3.

Pham, D.T., and Karaboga, D., 2000. *Intelligent OptimisationTechnique,*Springer, London, Great Britain, 337p.

Rothlauf, F., 2006.*Representations for Genetic and Evolutionary Algorithms*, Springer, Netherlands, 314p.

Samaher, H. A., 2007. *Object Oriented Classification of Forest Images UsingSoft Computing Approach,* 4rth International Conference: Sciences of Electronic,Technologies of Information and Telecommunications March 25-29, 2007 – TUNISIA.

Tan, f., X. Fu, et al.,2008.*A Genetic Algorithm-Based Method for Feature Subset Selection.*Soft Compute, p.p 111-120.

Waske, B. and Benediktsson, J.A., 2007. *Fusion of Support Vector Machines for Classification of MultisensorData,* IEEE Transactions on Geoscience and Remote Sensing, 45(12-1), December 2007.

Yang, M.S., and Wu, K.L., 2001.*A New Validity Index for Fuzzy Clustering*, IEEE International Fuzzy Systems Conference, pp. 89-92.

Yang, Y. F., 2006.*Genetic Algorithms for the Unsupervised Classification of Satellite Images,*Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Nienburger Str. 1, D-30167 Hannover, Germany.

Zhou, M., Shu, J. and Chen, Z., 2010.*Classification of hyperspetral remote sensing image based on Genetic Algorithm and SVM,Remote Sensing & Modeling of Ecosystems for Sustainability VII*, Proc. of SPIE Vol.7809, 78090A, doi:10.1117/12.860153.