AN INTRODUCTION TO BIOLOGICALLY INSPIRED ENGINEERING: GENETIC ALGORITHMS, EVOLUTIONARY COMPUTING AND ARTIFICIAL IMMUNE SYSTEMS – THEORY & APPLICATIONS IN CLASSIFYING REMOTE SENSING SATELLITE IMAGE DATA

Sharath Tadepalli (tadepals@purdue.edu), Graduate student – Geomatics area, School of Civil Engineering, Applied Statistics & Spatial Data analysis area, School of Statistics, Purdue University, West Lafayette, IN 47907, USA

ABSTRACT

Clustering and classification are the most significant aspects in the analysis of remote sensing data from an end user point of view. Biologically inspired engineering techniques have been adopted for some time now to achieve these objectives and compare their performance against existing mathematical and statistical techniques. Some of the common methods used in this regard are neural networks, artificial intelligence and neural fuzzy models against methods like maximum likelihood, machine learning and data mining from a statistical pattern recognition approach. This paper talks about the applicability and relevance of two other such biologically inspired engineering techniques for the spatial data classification problem - Genetic algorithms and Artificial immune systems. A theoretical introduction to GAs and AIS will be presented to the readers and their relevance tested in the spatial domain using multi-channel multi-class multi-spectral satellite imagery. A comparison would then be drawn on the performance of GAs against that of AIS. A comparison would also be drawn between evolutionary computing (GAs + AIS) and other biologically inspired methods including neural networks and neural fuzzy systems and traditional engineering methods borrowed from machine learning, data mining and support vector machines. Home brew remote sensing software, MultiSpec and genetic algorithm software GOSET are employed for producing the final thematic classification maps and the optimization results respectively. It is intended to introduce to the readers the enormous research potential that evolutionary computing techniques present and their great adaptability to the image classification endeavor. It has been noted that not enough research has been done in this regard to validate or disprove the claim that GAs and AIS outperform existing traditional methods and this paper is thus an endeavor to present to the scientific community results obtained from employing these approaches to assist the readers in appreciating their benefits. It will be demonstrated that extending the use of these two effective search procedures, the Genetic algorithm and the Artificial immune systems, leads to enhanced feature extraction, feature selection, parameter estimation and hence improved classification accuracies. It is shown that these methods are superior to the traditional statistical pattern recognition algorithms using overall end classification accuracy of the training data set and the Kappa coefficient reached, as performance measures. A relative comparison would then be drawn in terms of other performance specifications like model complexities (computational cost, time, order of processes), costs of misclassification and robustness of the methods.

Keywords

Genetic algorithms (GAs), Artificial immune systems (AIS), Biologically-inspired-Engineering (BIE), Evolutionary Computing (EC = GAs + AIS), clonal selection algorithm, spatial data, remote sensing etc

1. Introduction

Evolutionary computing techniques are techniques that ape the human genetic and immunological principles in their quest for search, optimization and learning. They differ from other BIE approaches like neural networks and neural fuzzy systems which involve both structural and parameter learning. This makes them trainable dynamical systems. GAs and AIS are essentially parameter learning models that are not trainable or dynamical. They differ from other classification algorithms used in the image processing, data mining and support vector

machine domains for being not error based structural risk minimization tools and suffer from serious mathematical and statistical inadequacies. However in this paper it is aimed to mimic the genetic and immunological principles for feature selection, extraction, data clustering and subsequent classification by symbolically pattern matching the aforementioned exercises within the human body to those from the spatial data set. A common rule base and end user requirement is what motivates this comparison and hence the endeavor.

1.1 What is geo-spatial remote sensing data?

[5] Spatial data is the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationships among such objects. Spatial data carries topological and or distance information and it is often organized by spatial indexing structures and accessed by spatial search methods. Some typical examples include: ground based spectrometer data, thermal sensor imagery, LIDAR imagery, RADAR imagery, MS/HS imagery, IR imagery, photogrammetric imagery and GPS data. Traditional (non-spatial) databases are concerned with only the attributes of objects. They make no explicit distinction between the location of an object and its other attributes. A given spatial database provides for the storage and manipulation of four aspects of its data – . Location component, . Topological component,. Attribute component and Metadata component The location component is a record of the position in geographical space that determines where something is and what form it takes. The topological component is a record of the logical relationships between different geographic objects. The attribute component is a record of the characteristics of things that determine what geographic objects represent and what properties they have. The metadata component is a thorough documentation of the contents of the overall database.

1.2 What are genetic algorithms and evolutionary methods?

[11] Genetic algorithms are search procedures based on the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with randomized yet structured information exchange with each new generation of strings inheriting only the desired properties from the preceding generation. In GAs a candidate solution is referred to as an individual that consists of strings of binary or real coded genes. As in the natural evolution process the parent individual generate offspring individuals by means of random variations. The resulting off spring solutions are evaluated for their effectives or fitness. Based on the rule of the survival of the fittest, less fit individuals are removed and this process of random variation and selection is repeated. More specifically genetic algorithms operate on a population of strings with the strings coded to represent some underlying parameter set. Reproduction, cross over, mutation and migration are applied to successive string populations to create new string populations. These operators are nothing more complex than random number generation, string copying and partial string exchanging. An illustration of gene (candidate) sequencing into a genome (mutated and reproduced off springs in the mating pool) is seen below:

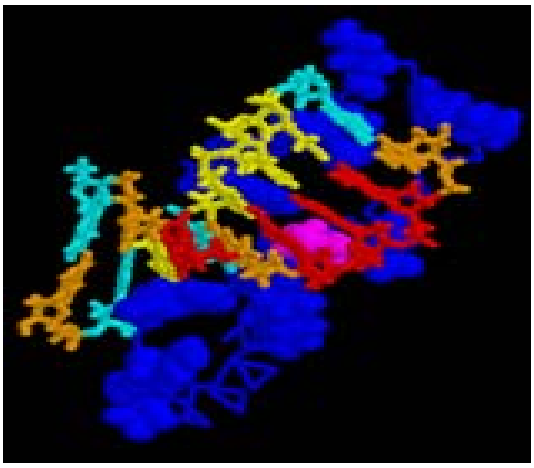


Figure 0

1.3 What are artificial immune systems?

The immune system is a complex and intricate system, perhaps in the same magnitude as that of the brain and the nervous systems. Just as in brain inspired synthetic neural networks, the natural immune system can be considered as a source of inspiration for developing intelligent methodologies towards problem solving with regard to its powerful information processing capabilities. The natural immune system is a complex system with several functional components. It employs a multilevel defense against invaders through non-specific (innate)

and specific (acquired) immune mechanisms. The main role of the immune system is to recognize all cells (or molecules) within the body and categorize those cells as self or non-self. The non-self cells are further categorized in order to stimulate an appropriate type of defensive mechanism. The immune system learns through evolution to distinguish between foreign *antigens* (e.g. bacteria, viruses, fungi, parasites, etc.) and the body's own cells or molecules. The human body maintains a large number of immune cells which circulate throughout the body. The *lymphocyte* as seen below is the main type of cell participating in the immune response that possesses the attributes of specificity, diversity, memory, and adaptivity. Other cells called *phogocytic* cells are accessory immune cells whose primary function is to provide facilities to eliminate antigens. There are two main types of lymphocytes, namely T cells and B cells. The primary lymphoid organs provide sites where lymphocytes mature (by undergoing training and testing) and become antigentically committed. One of the possible tests carried out is that of recognizing self and non-self cells. The lymphocytes that attack self-cells are immediately destroyed. T cells develop in the bone marrow but travel to the thymus to mature, whereas the B cells develop and mature in the bone marrow. The secondary lymphoid organs function to capture antigen and to provide sites where lymphocytes interact with the antigen to stimulate an immune response.

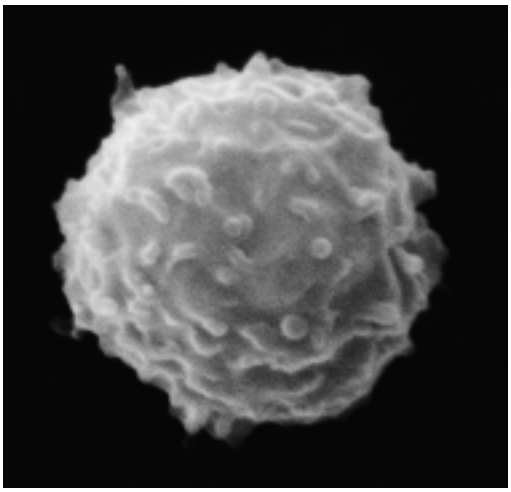


Figure 1

When an antigen invades the body, only a few of these immune cells can recognize the invader's peptides (distinguishing features). This recognition stimulates proliferation and differentiation of the cells that produce matching clones or (antibody). This process, called clonal expansion, generates a large population of antibody producing cells that are specific to the antigen. The clonal expansion of immune cells results in destroying or neutralizing the antigen.

1.4 Why biologically inspired engineering in remote sensing?

Evolutionary computing techniques differ from traditional methods given that they posses unique advantages like modeling missing data sets well that otherwise require EM algorithm, search the entire feasible solution space hence avoiding getting stuck at local minima, are not data driven or model dependent, inherit properties that are acceptable over evolutions assuring unique data dimensionality reduction, some what analogous to the PCA/ICA and work with very effective memory management & pattern matching skills. They operate on encoding of the parameter values and not necessarily the actual parameter values, use only the fitness or effectives ness values based on the objective function defined by the statistical distance metric and do not require derivative information or other collateral knowledge. They are probabilistic computations and not deterministic and are very efficient in handling problems with discrete search spaces. They are all well suited to the spatial image classification problem. Further we can loosely interpret the proliferation process in AIS & migration operation in GAs as being analogous to the feature selection process, the mutation process in both GAs and AIS as the feature extraction process, the cloning operation in AIS & the selective crossover operation in GAs as the clustering exercise which are the three main components of the problem at hand. Some of the other key features of the genome/immune system which provide several important aspects to the field of information processing may be summarized under the following terms of computation: Feature recognition, Anomaly (outlier) detection, Learning and Memory management (adaptability), Distributed detection, Threshold mechanism and Self-regulation.

2. The Data set:

The data obtained was a multi-spectral RGB color coded image of an agricultural land mass in north central Indiana. The data was recorded and stored as flight line FlC1.lan.

Table 1:
Description information for -- 'FLC1.LAN'

File format:	Erdas73
Image type:	Multispectral
Band interleave format:	BIL
Signed data:	No
Number of lines:	949
Number of columns:	220
Number of channels:	12
Number of bytes:	1
Number of bits:	8
Number of header bytes:	128
Number of pre-line bytes:	0
Number of post-line bytes:	0
Number of pre-channel bytes:	0
Number of post-channel bytes:	0
Line start:	1
Column start:	1

Figures 2 and 3 below depict the original data set and the digitized set respectively.

Figure 2

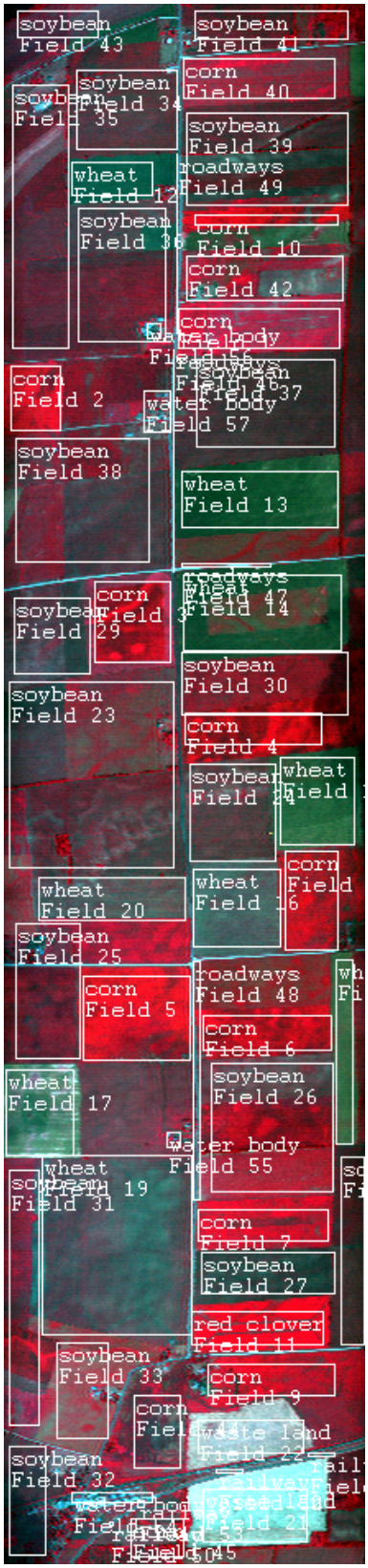


Figure 3

3. Related Work:

[1] Brandt C. K. Tso and Paul M. Mathersug recommended that with the increasing availability of multisource remotely sensed data sets, random field models, especially Markov random fields (MRF), have been found to provide a theoretically robust yet mathematical tractable way of coding multisource information and of modeling contextual behavior. It is well known that the performance of a model is dependent both on its functional form (in this case, the classification algorithm) and on the accuracy of the estimates of model parameters. In dealing with multisource data, the determination of source weighting and MRF model parameters is a difficult issue. They extend the methodology proposed by demonstrating that the use of an effective search procedure, the Genetic Algorithm, leads to improved parameter estimation and hence higher classification accuracies. [2] R. Khedam and A. Belhadj-Aissaa studied that in particular of remotely sensed satellite imagery, adjacent pixels are related or correlated, both because imaging sensors acquire significant portions of energy from adjacent pixels and because ground cover types generally occur over a region that is large compared with the size of a pixel. It seems clear that information from neighboring pixels should increase the discrimination capabilities of the pixel-based measured data, and thus, improve the classification accuracy and the interpretation efficiency. This information is referred to as the spatial contextual information. In recent years, many researchers have proven that the best methodological framework which allows integrating spatial contextual information in images classification is Markov Random Fields (MRF). In this paper, the authors present a contextual classification method based on a *maximum a posterior* (MAP) approach and MRF. An optimization problem arises and it will solved by using an optimization algorithm such as Iterated Conditional Modes (ICM) which occurs the definition and the control of some critical parameters: neighboring size, regularization parameter value and criterion convergence. [3] Liangpei Zhang, Yanfei Zhong and Pingxiang Li suggested that as a novel branch of computational intelligence, AIS has strong capabilities of pattern recognition, learning and associative memory, hence it is natural to view AIS as a powerful information processing and problem-solving paradigm in both the scientific and engineering fields.

Artificial immune systems possess nonlinear classification properties along with the biological properties such as self/nonself identification, positive and negative selection, and clonal selection. Therefore, AIS, like genetic algorithms and neural nets, is a tool for adaptive pattern recognition. However, few papers concern applications of AIS in feature extraction/classification of aerial or high resolution satellite image and how to apply it to remote sensing imagery classification is very difficult because of its characteristics of huge volume data. Remote sensing imagery classification task by artificial immune system is attempted and the preliminary results are provided. The classification task employs the property of clonal selection of immune system. The clonal selection proposes a description of the way the immune systems copes with the pathogens to mount an adaptive immune response. [4] John J. Szymanski,.et al in their paper, report on work using genetic programming to perform feature extraction simultaneously from multispectral and digital elevation model (DEM) data. They use the GENetic Imagery Exploitation (GENIE) software for this purpose, which produces image-processing software that inherently combines spatial and spectral processing. GENIE is particularly useful in exploratory studies of imagery, such as one often does in combining data from multiple sources. The user trains the software by painting the feature of interest with a simple graphical user interface. GENIE then uses genetic programming techniques to produce an image-processing pipeline. The authors demonstrate evolution of image processing algorithms that extract a range of land cover features including towns, wildfire burnscars, and forest.

4. The genetic algorithm model:

Early theories of inheritance proposed by ancient Greeks and medieval Europeans suggested that particles from all parts of the body form eggs & sperms and changes made in various parts of the body during an organism's life could be passed on to the next generation. The *gene* was thus conceptualized as a miniature human being in an incubator. Early theories of inheritance proposed that all genetic traits are inherited from the mother while some proposed a father-centric inheritance; the *blending theory* suggested during the 17^{th} century assured that both the egg and sperm contribute equally. Pioneering work in this area was initially taken up by Mendel whose laws of inheritance, incomplete dominance and independent segregation revealed that hereditary characteristics always occurred in pairs such that only one member of the pair is used in a *gamete*. Dominance and recessiveness were introduced for the fist time using ratios of various mating combinations and the word *gene* coined to indicate the physical and functional hereditary characteristic unit. The mechanism for dominance worked on the principle that genes result in the production of *enzymes* and for complete dominance one *allele* or the alternate form of a gene produces enough to achieve the desired effect. The mutation and evolution process of these alleles resulted in a *genotype* or the genetic makeup. These genotypes described the characteristic behaviors and traits of an individual called the *phenotype*. A typical mutation rate of a given gene was found to be 1 in 10^5 generations and since there were 10^4 genes per cell, mutation was pretty common. The genetic characteristics were carried through the *chromosomes* which were found to be mostly occurring in pairs and each human cell consisted of 23 such pairs. These cells then underwent mitosis or meiosis cell division processes for the reproduction stage to be activated. Man and animals higher up in the food chain were classified as *Eukaryotes* which had DNA organized into the chromosomes. Humans were called *haploids* following 3.4×10^6 base pairs of the DNA hairs in their system with an error rate of replication being very low. The message of the DNA was then transcribed into a RNA that translated to amino acids and proteins which resulted in human behavior and structural form. The genetic evolution process from a physical and engineering point of view dealt with two parent genes randomly selected from the mating pool of a given population of chromosomes, blending/mutating and then translating only the desired properties to their offsprings over subsequent generations. Consequently each pixel in the image space was interpreted as individual gene candidates and vice versa, given the spatial data set. The gene selection principle as described above is illustrated using the dendrogram plot below:

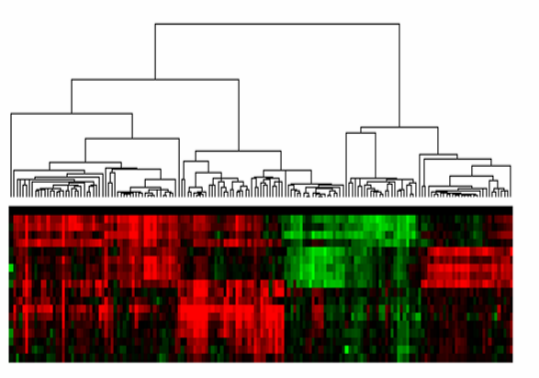


Figure 4

The gene selection algorithm:

Step 1. Initialize the population (pixels)

Step 2. Evaluate the fitness of each individual candidate (using a pre-defined distance metric)

Step 3. Select those candidates that are above the threshold value of the metric using roulette wheel or tournament selection to form a mating pool for reproduction

Step 4. Migrate or re-substitute the individuals with least affinity

Step 5. Protect and hold on to the individuals with the best affinity (elitism)

Step 4: Mate and randomly cross over those individuals that have the highest affinity

Step 6. Mutate these individuals now to create a new population

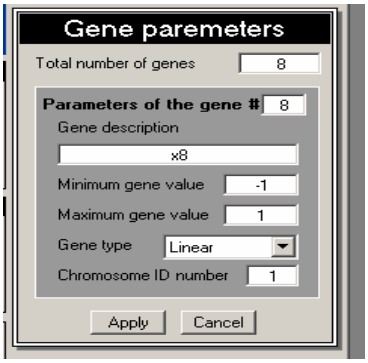
Step 7. Reevaluate the fitness of the new population candidates - check for diversity control with small fitness weights for individuals with numerous closely packed neighbors and vice versa

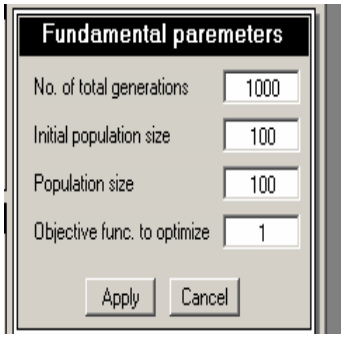
Step 8. Scale the fitness values to maintain the appropriate evolution pressure thorough out the evolution process

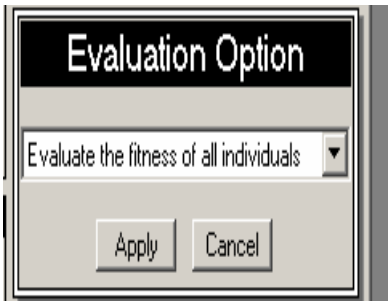
Step 9. Search the vicinity of the best individual for a better individual

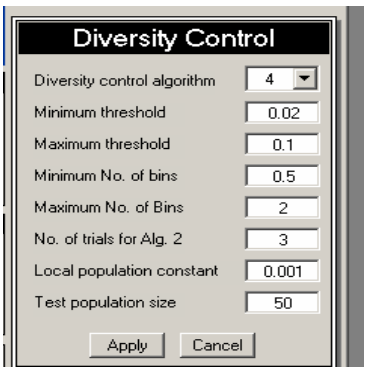
Step 10. Repeat - go back to step 2

The iterative computation of the evolution process with the predefined parameters can be seen in the illustration below.









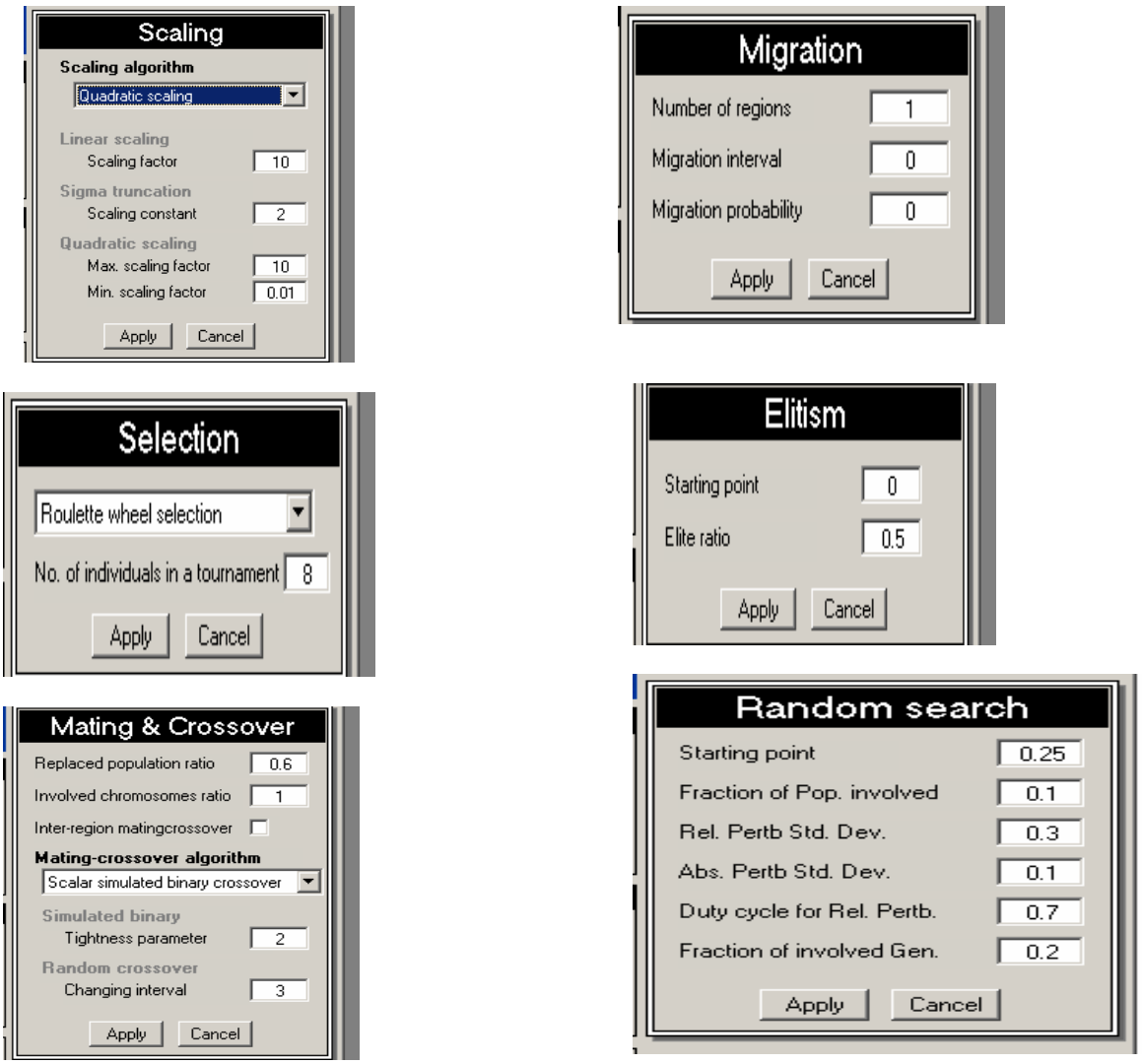


Figure 5

The entire solution space was spanned by the algorithm and individuals categorized into eligible features given the two constraints governed by the pre-defined objective functions 1 and 2 as seen below. This was the feature selection step. All individual candidate solutions contained in the hypercube [1 1] above a pre-defined threshold value of 0.4 were declared as those with high affinity. The individuals on the solution set corresponding to the Pareto front or boundary seen below were declared statistically significant candidates which were then approved for the feature extraction process.

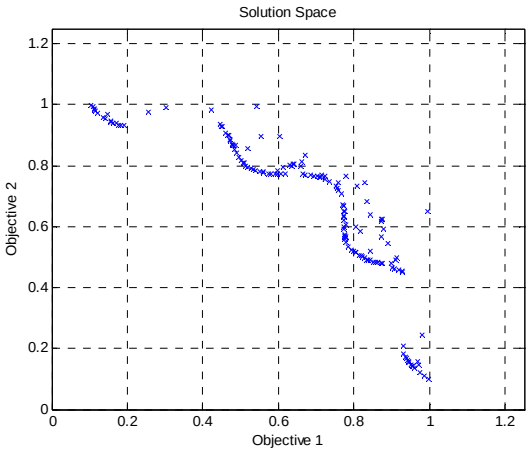


Figure 6

Best individual

Gene No.	Description	Value
1	x1	1
2	x2	1
3	x3	0.00034025
4	x4	-0.76864
5	x5	0.18155
6	x6	-0.19449
7	x7	0.34874
8	x8	-0.9022

Table 2

As seen from the table above and the figure below, eight different classes were extracted and their corresponding fitness values were evaluated using all the selected pixels against a given threshold value. The green '+' indicate all the pixels that satisfied the given criteria that were then mutated to the clustering process while the red 'o' were the statistically insignificant ones that failed the threshold test and were hence dropped from the population pool.

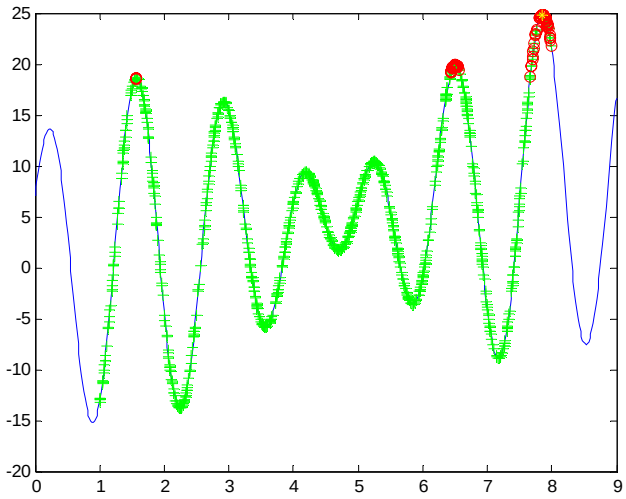


Figure 7

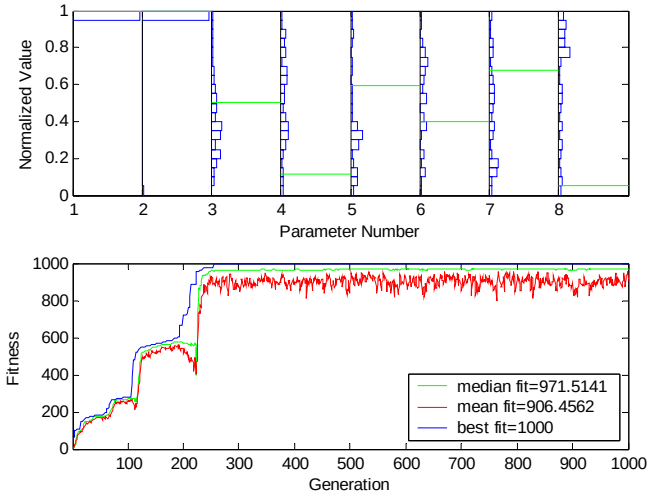


Figure 8

Finally the clustering of the selected individual candidate solution (pixels) was done using the mutated, reproduced, migrated, elite, scaled and diversified points from the final selection pool. The clustering process confirmed to the initial assumption on reconnaissance that eight different clusters exist in the data set. This can be concluded from the plot above which also illustrates that most of the data points were essentially Gaussian distributed as seen from the histogram bins. The entire evolution process was slow and tedious leading to over 300 generations before convergence was reached.

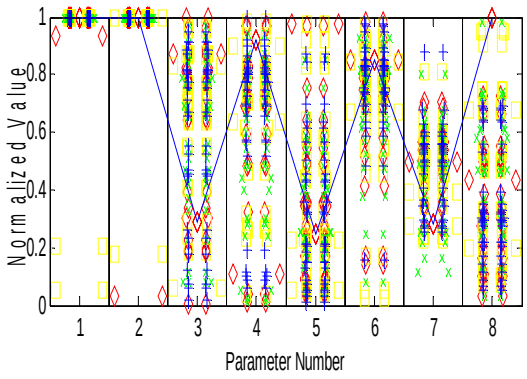


Figure 9

We observe that most of the data points were clustered in bins - 3, 4, 5 and 6 that correspond to corn, soybean, wheat and red clover content of the selected data set. This validated our initial assumption that most of the data set falls under vegetative land use and land cover and is essentially agricultural land with the afore-mentioned crops dominating the cultivation pattern.

Statistics for generation 1000
Best fitness = 1000
Mean fitness = 906.4562
Median fitness = 971.5141
Number of evaluations = 101630

Absolute computation times for generation 1000
OWV: 0.00e+000 DC: 3.10e-002 SCALE: 0.00e+000
SELECT: 1.60e-002 MC: 3.10e-002 MUT: 0.00e+000
MIGRATE: 0.00e+000 EVAL: 1.60e-002 ELITE: 1.60e-002
RS: 1.50e-002 STAT: 0.00e+000 REPORT: 2.03e-001

Relative computation times for generation 1000
OWV: 0.00 DC: 9.45 SCALE: 0.00
SELECT: 4.88 MC: 9.45 MUT: 0.00
MIGRATE: 0.00 EVAL: 4.88 ELITE: 4.88
RS: 4.57 STAT: 0.00 REPORT: 61.89

5. The artificial immune systems model:

The method of image sensing proposed utilizes properties of the adaptive immune system. This part of the immune system is made up mostly of leukocytes, or white blood cells. About 25 % of these cells are the specialized group known as lymphocytes. As discussed earlier, lymphocytes are divided into two groups: T cells and B cells. These two types of cells work together to recognize antigens, or unwanted invasions of the body. The way that they detect and then detain antigens makes them a useful model for data analysis. In the human body, the purpose of the B and T cells is to identify and suppress any antigen that enters the body. Each B cell can produce unique antibodies, and in the human body these antibodies are capable of detecting about 1 million different antigens. When a B cell encounters an antigen that matches its antibody, it engulfs the antigen and partially digests it. The B cell then displays a protein indicator on its surface. When the level of the indicator proteins exceeds a threshold, T helper cells trigger the next response, which includes increased antibody reproduction and the beginning of B cell clonal selection. The clonal selection process is an asexual reproduction. Once the T helper cells recognize that a certain level of antigen has entered the body, it triggers mitosis of the appropriate B cell. These cells divide and undergo mutation so that each new cell produces a slightly different antibody. The cells continue dividing, and the ones whose antibodies most closely match the antigen (highest affinity) remain and while those with the lowest affinity die off. After a few generations, the cells mature and their division create plasma cells and memory cells. Plasma cells are large cells whose only purpose is to create antibodies so the antigens can be detected by macrophages and digested. Memory cells are B cells with a very high affinity for the specific antigen. They are stored in the body in order to mount a faster defense in future encounters with the same antigen. Since their antibodies will have a nearly perfect match for the antigen, memory cells begin producing plasma cells as soon as the antigen is detected, significantly decreasing response time. Corresponding to the image classification problem one antigen activated was then assigned to one pixel. The clonal selection principle described above is illustrated below:

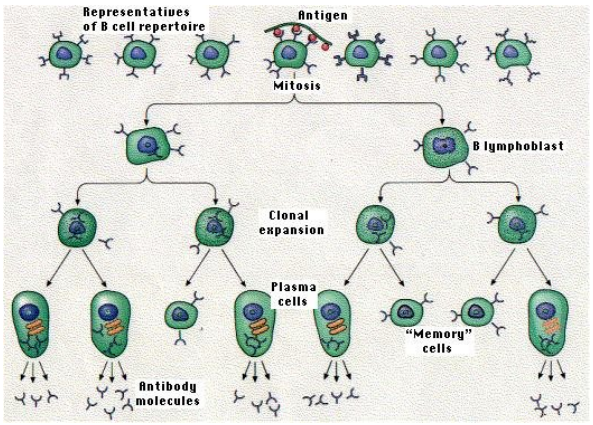


Figure 10

The clonal selection algorithm:

Step 1. Generate possible solutions

Step 2. Evaluate the fitness of each individual candidate (using a pre-defined distance metric)

Step 3. Select best individuals by highest affinity

Step 4. Clone 'n' best individuals to create a sample population

Step 5. Submit to hyper mutation and maturation to create a new population

Step 6. Reselect best individuals from new population to create a new solution set

Step 7. Supplement old solution set with the new ones

Step 8. Replace antibodies to introduce diversity control

Step 9. Operate on simulated annealing approach for better scaling and bias reduction

Step 10. Repeat – go back to step 2

As seen below in figure 11 below, eight different directions were initially assigned to the simulated annealing approach of the clonal selection algorithm with each direction corresponding to each of the eight independent clusters. The unsymmetrical plot indicates that the data points pertaining to majority class of corn, wheat, soybean and red clover are dominant enough to skew the weights towards themselves. Also the decision boundary hyper plane between these classes would hence be more regularized and well defined given greater uniformity within the respective clusters. This called for data standardization, normalization and scaling to offset the problem of skewed multi-classes through the simulated annealing process, which ensured better diversity control.

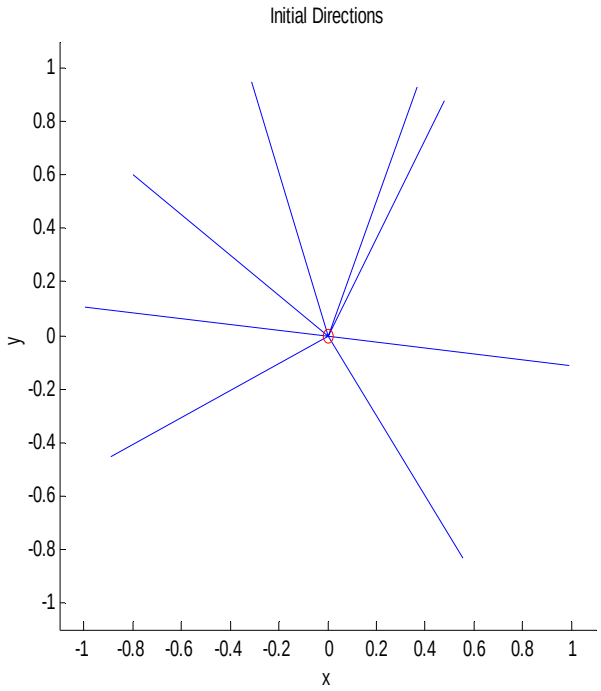


Figure 11

After the final feature selection process was completed the data along the principal component dimensions was retained the most, after the prior threshold test. This ensured that the clustering process would be as unbiased as possible along each direction or for each cluster i.e. to classify a new pixel the algorithm would treat all possible clusters with equal probability although the dimensionality reduction forced most of the statistically irrelevant data points from non-significant clusters to be eliminated. This is illustrated below.

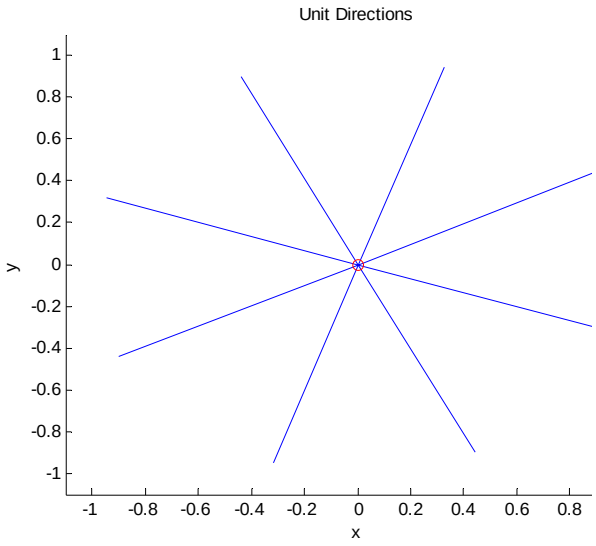


Figure 12

The feature extraction was then done based on the corresponding evolution energy or pressure corresponding to number of proliferated components produced. As seen below all proliferations that were extracted as features were based on the energy function value starting from 29 to 99 in steps of 10. As before samples pertaining to clusters of corn, wheat and soybean were the most dominant lying between the ranges of 50 to 90.

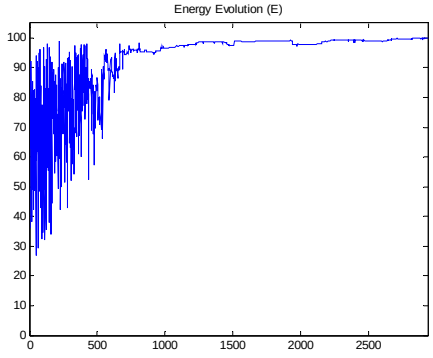


Figure 13

With feature extraction now complete the clustering operation was carried out following the cloning of the proliferated (extracted) features. Seen below is the result of the clustering process. The clusters corresponding to the majority class of corn, wheat, soybean and red clover line up at the peaks while those belonging to the minority class of road ways, railways, waste land and water body are at the troughs of the 3D grid surface.

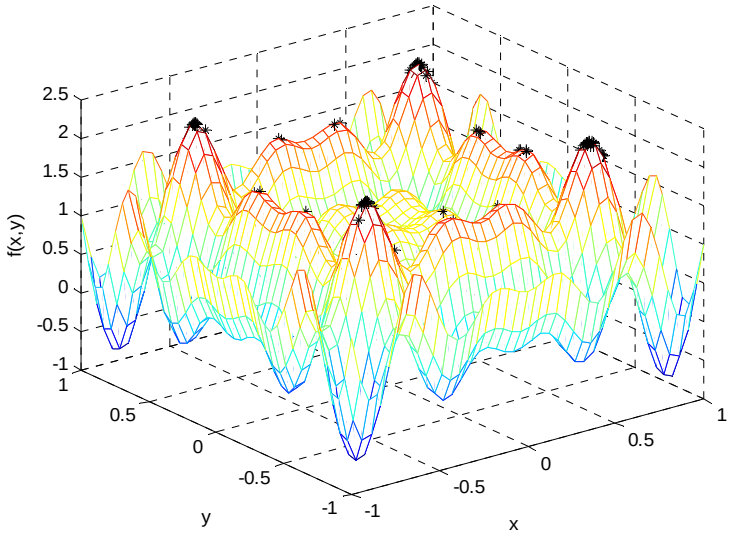


Figure 14

6. Experimental results:

The flowing table of accuracies depicts the performance of the listed methods on the same data set. It is seen that evolutionary computing techniques perform very well, better than classical statistical techniques but not up to the class of data mining, neural networks and support vector machines. The reasons for the discrepancies will be dwelt into later.

Table 3: Table of accuracies

Method	**Algorithms**	**Training class accuracy**
MultiSpec	FS,	~
	dbFE, isodata,	**86.40%/85.40 %**
	Bayes/ML	
Supervised	K-means, EM	~
Learning	and	**97.81%**
	Bayes	
Unsupervised	CART,	
Learning	bootstrapping,	
	Bayes	

Neural	Backprop,	~
Networks	AVQ,	**98.02%**
	Adaline,	
	SOM, LVQ	
Fuzzy Logic	Subclustering,	~
and Neural	fuzzy C –	**96.02%**
Fuzzy	means	
Systems		
Supervised	PAM, Bayes	~
Learning		**96.56%**
Binary and	Gap,	~
MSVM	multilayer	**98.43%**
	perceptron	
GAs	mutation, selection	~
	migration, elitism	**95.00%**
AIS	CSA, simulated	~
	annealing, mutation	**92.10%**
	proliferation	

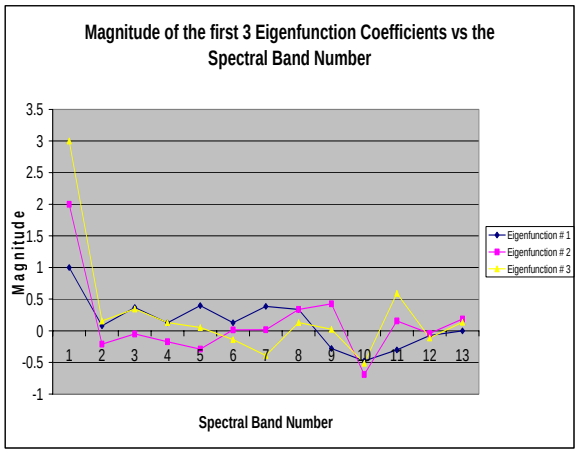


Figure 15

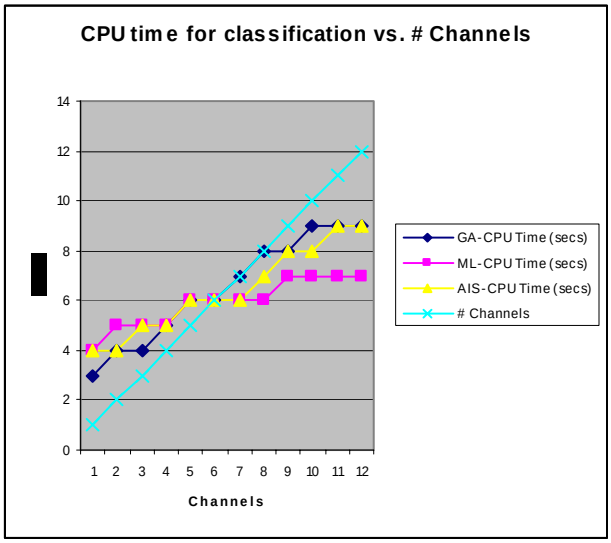


Figure 16

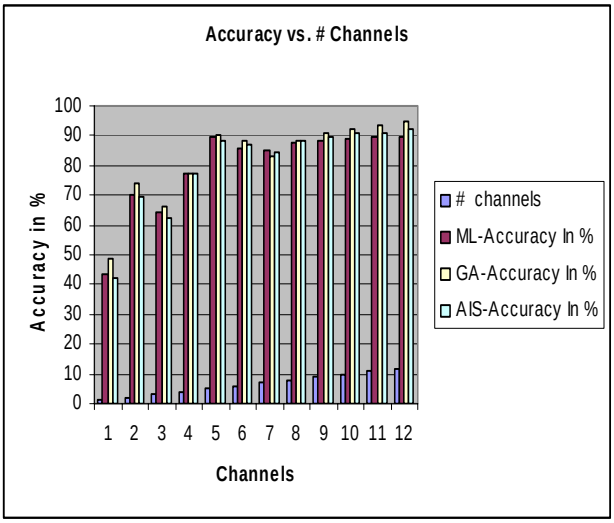


Figure 17

From the plot in figure 15 we see that considerable data dimensionality reduction was achieved using the AIS model which follows the yellow line and the GA model depicted using the magenta line while the classical model (PCA-ML) plotted using the blue line was not that significant.

6a. GAs vs. MultiSpec (Maximum Likelihood):

Table 4: Confusion Matrix (GA)

Project	Reference		Number of Samples in Thematic Image Class									
Class	Class	Accuracy+	Number	0	1	2	3	4	5	6	7	8
Name	Number	(%)	Samples	background	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
Class 1	1	96.0	1797	0	1726	0	70	0	0	0	0	1
Class 2	2	98.7	2938	0	467	289	430	495	392	289	386	190
Class 3	3	98.3	381	0	35	44	184	30	22	23	31	12
Class 4	4	88.1	26595	0	2	456	144	20777	134	792	1528	2762
Class 5	5	87.7	35182	0	39	3381	2948	793	16798	4788	1656	4779
Class 6	6	82.3	16724	0	5	281	885	242	4810	5395	2986	2120
Class 7	7	90.2	18708	0	10	699	203	3843	3728	2971	3750	3504
Class 8	8	98.4	44826	0	235	2092	1780	8988	6525	2522	5454	17230
TOTAL			147151	0	2519	7242	6644	35168	32409	16780	15791	30598
Reliability Accuracy (%)					68.5	4	2.8	59.1	51.8	32.2	23.7	56.3
OVERALL CLASS PERFORMANCE (139793 / 147151) = 94.99%												
Kappa Statistic (X100) = 52.3%. Kappa Variance = 0.000002												

9 CPU seconds for classification

Table 5: Confusion Matrix (ML)

Project	Reference		Number of Samples in Class								
Class	Class	Accuracy+	Number	1	2	3	4	5	6	7	8
Name	Number	(%)	Samples	corn	red clover	wheat	waste land	soybean	Road ways	Rail way	water body
corn	1	90.0	25814	18058	4507	1319	5	1026	439	321	105
red	2	84.1	1701	211	1431	2	0	15	24	13	5

clover											
wheat	3	89.8	30064	597	59	26996	65	1071	414	602	215
waste land	4	98.4	3477	4	2	1	3421	1	6	20	7
soybean	5	92.3	72484	15113	1082	15789	44	37895	845	1021	561
roadways	6	80.3	1141	136	27	53	11	48	574	211	74
railway	7	80.1	136	3	0	8	1	4	7	109	4
water body	8	82.7	972	115	7	84	4	44	130	234	318
TOTAL			135789	34237	7115	44252	3551	40104	2439	2531	1289
Reliability Accuracy (%)				52.7	20.1	61	96.3	94.5	23.5	4.3	24.7
		OVERALL CLASS PERFORMANCE (115964 / 135789) = 85.4%									
		Kappa Statistic (X100) = 32.0%. Kappa Variance = 0.000003									

7 CPU seconds for classification

With reference to figures 18 and 19 below, we observe smoother and much finer texture to the end classification thematic map of the data set following the GA model in comparison to that from the maximum likelihood model as seen in the figures below. This is attributed to the enhanced statistical feature selection and extraction that ensured a better and more uniform neighborhood for classification. The time taken for the end result was slower in the GA domain understandably due to the sampling and resampling process following mutation and reproduction being tedious and cumbersome. However, from figures 16 and 17 above, a vast improvement in the end classification accuracy is noticed. The Kappa statistic also increased substantially confirming that the GA model was much better suited to the spatial data classification problem than the ML model.

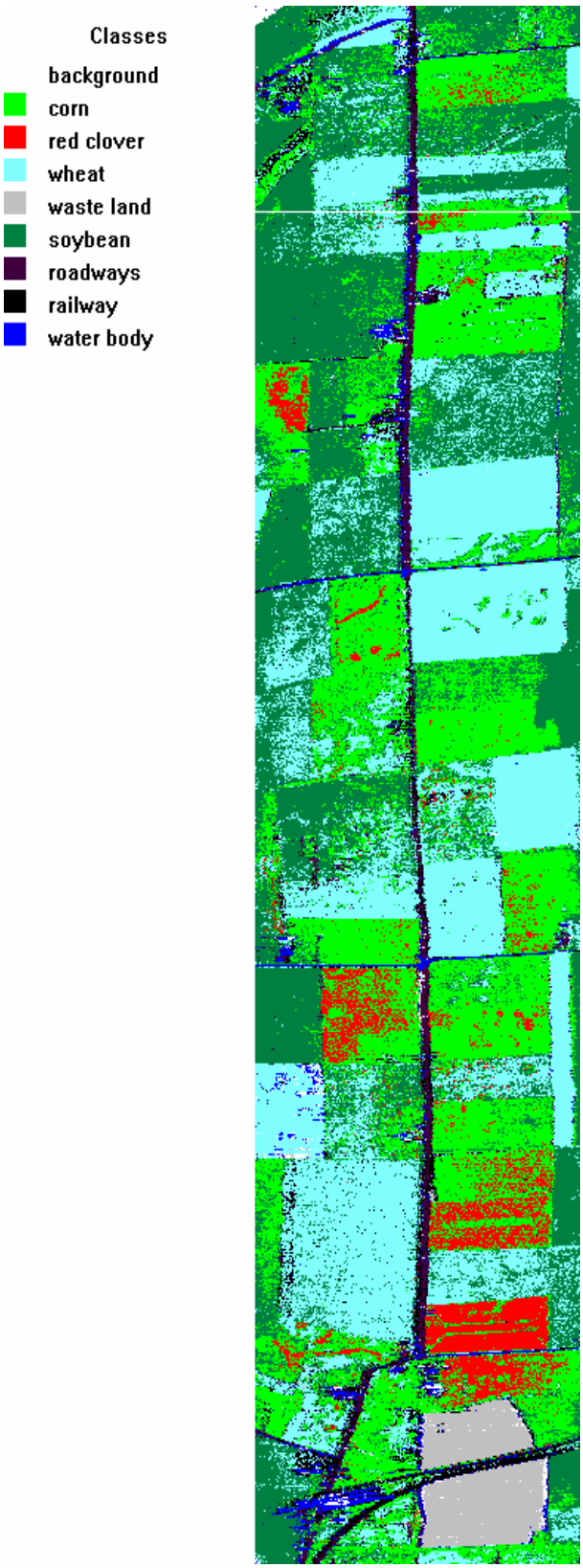


Figure 18 – ML thematic map

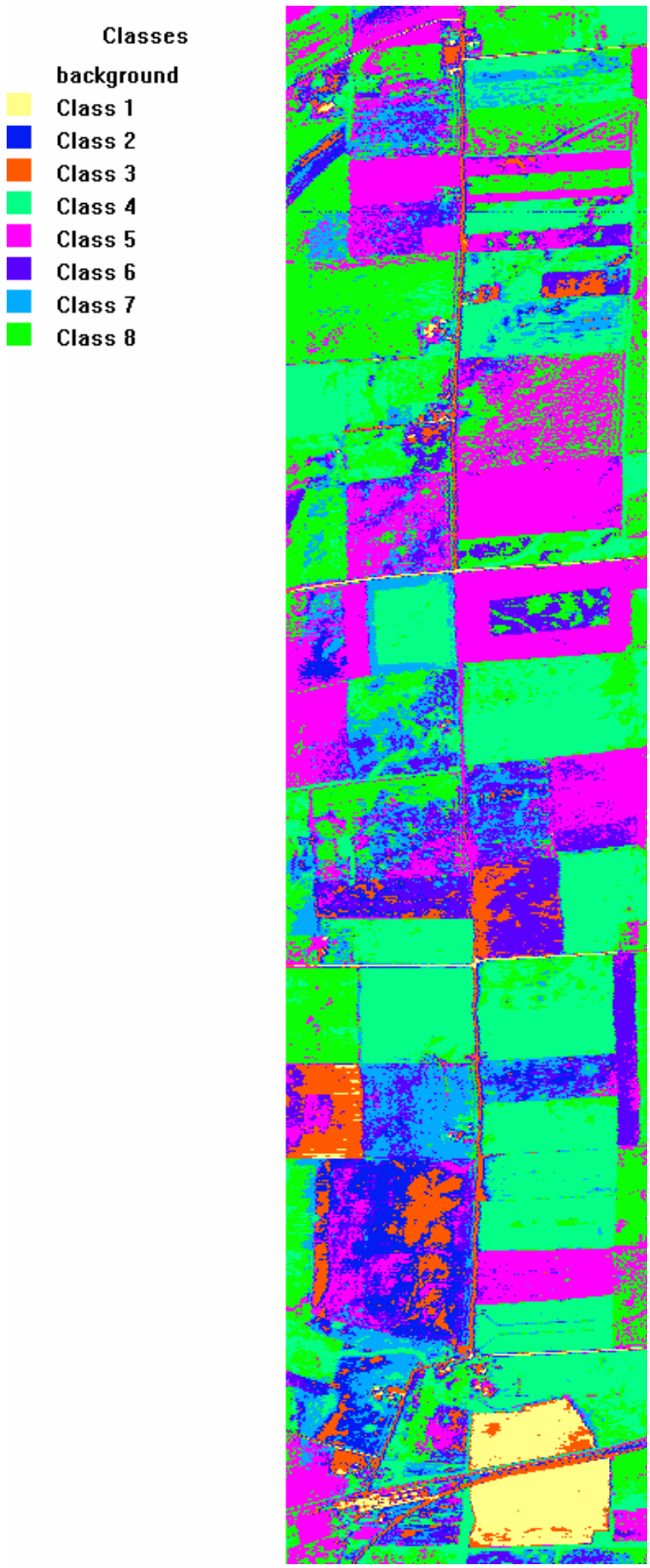


Figure 19 – GA thematic map

6b. AIS vs. MultiSpec (Maximum Likelihood):

Table 6: Confusion Matrix (AIS)

Project	Reference		Number of Samples in Thematic Image Class									
Class	Class	Accuracy+	Number	0	1	2	3	4	5	6	7	8
Name	Number	(%)	Samples	background	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
Class 1	1	97.5	1797	5	1788	0	3	0	0	0	0	1
Class 2	2	93.3	2938	1033	14	392	343	316	57	232	258	293
Class 3	3	95.2	381	116	19	27	96	28	4	52	23	16
Class 4	4	80.2	26595	4604	6	170	108	16001	312	882	1159	3353
Class 5	5	97.6	35182	978	1212	8538	10862	698	2689	3891	5128	1186
Class 6	6	83.8	16724	704	244	1751	3657	175	1180	3981	3478	1554
Class 7	7	89.2	18708	499	2	4426	344	3525	589	2466	3597	3260
Class 8	8	91.8	44826	2767	52	4079	2557	7731	1427	3636	8308	14269
TOTAL			147151	10706	3337	19383	17970	28474	6258	15140	21951	23932
Reliability Accuracy (%)					53.6	2	0.5	56.2	43	26.3	16.4	59.6
		OVERALL CLASS PERFORMANCE (135526 / 147151) = 92.1%										
		Kappa Statistic (X100) = 48.6%. Kappa Variance = 0.000002										

9 CPU seconds for classification

Similar results were seen between the AIS model and the GA model, both when compared individually to the ML model. Once again there was significant improvement in the classification accuracy and the Kappa statistic but AIS was slower than the classical model following an elaborate cloning, maturation and proliferation operation as seen from figures 16 and 17 above. The final thematic classification maps produced were smoother than those from the ML domain but coarser than those from the GA model as inferred from figure 20 below. Reasons for these were analyzed and summarized subsequently.

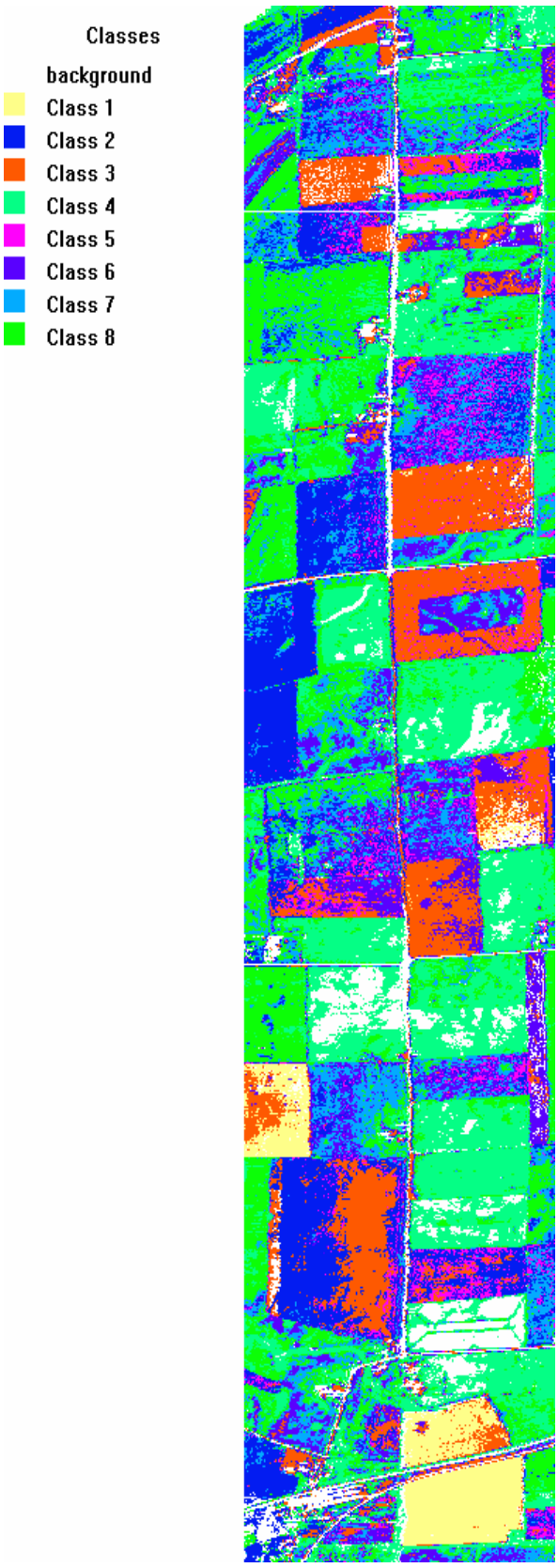


Figure 20 – AIS thematic map

6c. GAs vs. AIS:

As noted from tables 4 and 6 above, genetic algorithms performed slightly better than artificial immune systems. This can be attributed to the probabilistic nature of GAs being higher than that of AIS which induces greater randomness to the crossover – mutation – migration - search and selection processes that can be tricky and problematic given pixels that adjoin two clusters. When we talk about such pixels that are very closely located to the boundary in between two clusters, the statistical threshold values have to be precisely defined. This property ensures that outliers are better weighted and noisy data (pixels on the borders) better analyzed and hence GAs being slightly more probabilistic ensured better classification. AIS does not operate on migration and elitism but clones and proliferates systematically and hence is slightly more deterministic per pixel when compared to GAs operating on the same threshold. This phenomena when generalized to the entire data set resulted in a slightly better feature selection and feature extraction in the GA domain and hence a slightly better end classification accuracy. Further the AIS model was generated using a simulated annealing approach of the clonal selection algorithm which assigned uniform data direction to each cluster which resulted in a substantial data dimensionality reduction. The same was achieved through selective cross over and mutation operations spread over 300 generations in the GA network. As seen from figure 15 above, chances of loss of statistically significant pixels could be mathematically viewed to be slightly more in the former case than the latter which might be another reason for the difference in the classification accuracies.

7. Evolutionary computing vs. Classical methods:

As seen from table 3 above, evolutionary computing techniques outperformed classical statistical techniques like Bayesian and maximum likelihood classification because of superior prior feature selection and feature extraction procedures. The statistical distance metrics employed were very similar in both the techniques but pixel selection was better in the EC domain given that cloning in the AIS model and mutation-selection in the GA model ensured the pixels chosen were statistically significant

and the right pixel neighborhood was employed hence due to the iterative (evolutionary) computation which was missing in the Bayes/ML methods. However EC was not as accurate as the supervised and unsupervised machine learning & data mining and support vector machines based algorithms given the fact there was no way of back propagating the error of misclassification within the EC domain. Extensive mathematical and statistical analysis followed with structural and empirical risk minimization enabled superior outlier detection, feature selection and feature extraction and thorough data cleaning through data standardization, normalization and scaling which resulted in a more precise classification within these methods. However EC techniques have the unique advantage similar to SVMs for being not data driven or model based and hence are more readily adaptable.

8. Evolutionary computing vs. Other BIE techniques:

With reference to table 3 above, we note that evolutionary computing techniques did not match up to the performance of neural networks and neural fuzzy systems due to the advantages inherent to the latter models given the residual driven error based structural and parameter learning/estimations of the neural network model and a precise learning centered IF-THEN-ELSE rule base with error analysis in the neural fuzzy domain. These features enabled these systems to adapt to the sptatio-temporal nature of the data set, data irregularities and other data based problems including non-uniform, non-normal, correlated, factor interactive and non-linear format of the given data set better than the EC algorithms which resulted in a better end classification accuracy.

9. Conclusions

This paper presents an alternate method of evolutionary computing using genetic algorithms and artificial immune systems to address the spatial data classification problem given the multispectral remote sensing satellite imagery. It has been shown that these techniques perform better that some of the conventional and classical statistics based techniques previously employed but lag behind some of the new techniques being used since recent times. They posses unique advantages like modeling missing data sets well that otherwise require EM algorithm, search the entire feasible solution space and hence avoid getting stuck at local minima, are not data driven or model dependent, inherit properties that are acceptable over evolutions assuring unique data dimensionality reduction some what analogous to the PCA/ICA and work with very effective memory management & pattern matching skills. Evolutionary computing though does present a huge research potential and given proper manipulation to these algorithms, will definitely match up to the performance of some of the other better known more widely used models. Some of the distinct advantages that existing methods possess include greater robustness and adaptability to the seasonal and time varying nature of the spatial-temporal data set, greater generalization performance, error metric centered analysis and better data management skills. However as explained earlier EC algorithms are data independent and hence easily and readily applied. Model complexities and constraints originating from cost, time and order of processes is another issue with EC techniques. It is always subjective with a huge degree of uncertainty on what a pixel means within the GA and AIS networks. For example, in this research a pixel resembles a gene candidate in the GA stream and an antigen in the AIS stream. Replacing this ideology with a 4x4 or and 8x8 neighborhood to a candidate and antigen might yield better or worse results but the process might be much faster and less expensive. These are some of the factors that can be looked at from a research point of view. Further can cost function, error back propagation and enhanced statistical measures be incorporated into the EC system? Can the classification be extended to other types of imagery? These are some questions that pose a good challenge and warranty a great follow up to this study. In summary artificial immune systems or immunological computation and genetic algorithms (sometimes referred to independently as evolutionary computation) are relatively new fields with lots of research potential. There are currently a few algorithms that can be used to model and mimic the human genome/immune system. However there is a lot about these systems that are still not clearly interpreted. Arguably, with a better and complete understanding of these structures, suitable models that simulate various components of the genome/immune system can be created.

Future Work

[8] It is proposed to extend this study to the analysis and classification of hyper-spectral, thermal, infrared, RADAR and LIDAR imagery so as to achieve maximum and optimal data classification and information extraction. Effect of these image (data) collection methods on the algorithms employed and vice-versa will be investigated. For example, thermal and IR imagery are collected in the middle & far IR and thermal regions of the EM spectrum while RADAR and LIDAR data correspond to the microwave & laser regions. Effects and properties of these regions from a physical and statistical standpoint are of keen interest with respect to final data classification. Also, a query based spatial data base development using Geographic Information Systems (GIS) technology, integrating with Global Positioning Systems (GPS) and bi-static RADAR sources to determine the effect of Z-dimension, time series based statistical modeling & analysis and applying statistical space inference methods would be interesting follows on the project given the data set. In conclusion, archiving the results obtained from using MultiSpec, neural networks, fuzzy logic systems, neural fuzzy systems, machine learning & data mining, evolutionary computing and SVM techniques on the same data set is possible. Testing the relevance and applicability of other useful methods from the areas of digital image processing (MAP segmentation, EM clustering, content based feature extraction, MRF modeling with ICM optimization) and scientific data visualization for analysis is noteworthy and warranties attention from a research standpoint.

Acknowledgements:

I am grateful to Dr. David Landgrebe, Professor, Electrical and Computer Engineering, Purdue University for providing me with the data set and access to the MultiSpec software for the analysis. I am thankful to Dr. Okan Ersoy, Professor, Electrical and Computer Engineering, Purdue University, Dr. Carla Brodley, Professor, Electrical and Computer Engineering, Purdue University, Dr. Ragu Balakrishnan, Professor, Electrical and Computer Engineering, Purdue University, and Dr. C.S. George Lee, Professor, Electrical and Computer Engineering, Purdue University for their guidance and useful comments in organizing the content of this paper. Further, much help has also been received from their class notes with regards to neural networks, fuzzy logic & neural fuzzy systems, artificial immune systems and machine learning & data mining. Special thanks to Dr. Scott Sudhoff for providing me with the GOSET software and his class lecture notes on genetic algorithms and evolutionary computing. Thanks are also due to Dr. William Cleveland, Professor, Department of Statistics, Purdue University, and Dr. Jongwoo Song, Professor, Department of Statistics, Purdue University, for their encouragement and support through journals, papers and class material that helped realize this work. I am also appreciative of my peers for their reviews on the study and their constructive suggestions for the betterment of the research and the presentation of the final paper.

Bibliography and References:

[1] *Brandt C. K. Tso and Paul M. Mather*-- Classification of Multisource Remote Sensing Imagery using a Genetic Algorithm and Markov Random Fields, IEEE Transactions on Geosciences and Remote Sensing, Vol. 37, No. 3, May 1999

[2] *R. Khedama, A. Belhadj-Aissaa* -- Contextual Classification of Remotely Sensed Data using MAP Approach and MRF, Image Processing Laboratory, Electronic and Computer Science Faculty, Technology and Sciences University

[3] *Liangpei Zhang, Yanfei Zhong, Pingxiang Li* -- Applications of Artificial Immune Systems in Remote Sensing Image Classification, State Key Laboratory of Information Engineering in Surveying Mapping & Remote Sensing, Wuhan University

[4] *John J. Szymanski, .et al* -- Feature Extraction from Multiple Data Sources Using Genetic Programming, Los Alamos National Laboratory

[5] *Sharath Tadepalli* -- Applications of Machine Learning, Knowledge Discovery and Data Mining in the Analysis, Classification and Assessment of Remotely Sensed Geospatial Satellite data, ASPRS 2004 Annual Conference

[6] *Sharath Tadepalli* -- Analysis and Classification of Remotely Sensed Satellite Data using Neural Networks and Fuzzy Logic Systems techniques, 2003 ACSM-APLS Conference and Technology Exhibition

[7] *Sharath Tadepalli* -- Exploration in High Dimensional Data

[8] *Sharath Tadepalli* -- A Comparative Evaluation on the Performance of Binary and

Multi-Class Support Vector Machines in the Analysis and Classification of Remotely Sensed Satellite Data, ASA-JSM 2005 Annual Conference
[9] *Leandro Nunes de Castro, Fernando J. Von Zuben* -- The Clonal Selection Algorithm with Engineering Applications
[10] *D. Dasgupta* -- Artificial Immune Systems and Their Application, Springer-Verlag, 1999
[11] *David E. Goldberg* -- Genetic Algorithms in Search, Optimization & Machine Learning, Pearson Education, 1999
[12] *Chuck Staben* -- Gene Structure and Identification, National Science Foundation
[13]DataSource: http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/documentation.html