Comparison of Feature Selection Techniques for SVM Classification

Gidudu Anthony[a] and Heinz Ruther[b]

[a] Dept. of Surveying, Makerere University, P. O. Box 7062, Kampala, Uganda - agidudu@tech.mak.ac.ug

[b] Dept. of Geomatics, University of Cape Town, 7701, Rondebosch, Cape Town, South Africa - hruther@ebe.uct.ac.za

Key Words: *Support Vector Machines, Feature Selection, Exhaustive Search, Population Based Incremental Learning*

Abstract:

The use of satellite imagery in the derivation of land cover information has yielded immense dividends to numerous application fields such as environmental monitoring and modeling, map making and revision and urban studies. The extraction of this information from images is made possible by various classification algorithms each with different advantages and disadvantages. Support Vector machines (SVMs) are a new classifier with roots in statistical learning theory and their success in fields like machine vision have drawn the attention of the remote sensing community. Previous studies have focused on how SVMs compare with traditional classifiers such as maximum likelihood and minimum distance to means classifiers. They have also been compared to newer generation classifiers such as decision trees and artificial neural networks. In this research the understanding of the application of SVMs to image classification is furthered by proposing feature selection as a way in which remote sensing data can be optimized. Feature selection involves selecting a subset of features (e.g. bands) from the original set of bands that captures the relevant properties of the data to enable adequate classification. Two feature selection techniques are explored namely exhaustive search and population based incremental learning. Of critical importance to any feature selection technique is the choice of criterion function. In this research a new criterion function called Thornton's separability index has been successfully deployed for the optimization of remote sensing data for SVM classification.

1.0 INTRODUCTION

Feature selection is defined as "the search for a subset of the original measurement features that provide an optimal trade off between probability error and cost of classification" (Swain and Davis, 1978). It involves selecting a subset from the original set of features (e.g. bands) that captures the relevant properties of the data (Gilad-Bachrach et al, 2004) to enable adequate classification (Wu and Linders, 2000). Feature selection is, in part, motivated by Hughes phenomenon (Hughes, 1968), which postulates that in the presence of a limited training sample size, contrary to intuition, the mean accuracy doesn't always increase with additional measurements. Rather it exhibits a peaking effect i.e. the accuracy will increase until a certain point beyond which increase in additional measurement will yield no significant improvement in classification accuracy (Muasher and Landgrebe, 1982). The challenge therefore is to identify the subset of bands which will yield similar, if not better, accuracies as compared to when all the bands are used in a classification task. In effect, feature selection is an optimization task. In this research two feature selection techniques namely exhaustive search and population based incremental learning (PBIL) are investigated in as far as their ability to optimize remote sensing data for Support Vector Machine (SVM) Classification.

2.0 SUPPORT VECTOR MACHINES

Support Vector Maachines (SVMs) are a new supervised classification technique that has its roots in Statistical Learning Theory (Vapnik, 1995). Having taken root in machine vision fields such as character, handwriting digit and text recognition (Vapnik, 1995; Joachims, 1998), there has been increased interest in their application to image classification (Huang et al, 2002; Mahesh and Mather, 2003). SVMs are non-parametric hence they boast the robustness associated with Artificial Neural Networks (Foody and Mathur, 2004a; Foody and Mathur, 2004b) and other nonparametric classifiers.

SVMs function by nonlinearly projecting the training data in the input space to a feature space of higher (infinite) dimension by use of a kernel function. This results in a linearly separable dataset that can be separated by a linear classifier. This process enables the classification of remote sensing datasets which are usually nonlinearly separable in the input space. In many instances, classification in high dimension feature spaces results in overfitting in the input space, however, in SVMs, overfitting is controlled through the principle of structural risk minimization (Vapnik, 1995). The empirical risk of misclassification is minimised by maximizing the margin between the data points and the decision boundary (Mashao, 2004). In practice this criterion is softened to the minimisation of a cost factor involving both the complexity of the classifier and the degree to which marginal points are misclassified, and the tradeoff between these factors is managed through a margin of error parameter (usually designated C) which is tuned through cross-validation procedures (Mashao, 2004). The functions used to project the data from input space to feature space are sometimes called kernels (or kernel machines), examples of which include polynomial, Gaussian (more commonly referred to as radial basis functions) and sigmoid functions. Each function has parameters which have to be determined prior to classification and they are usually determined through a cross validation process. A deeper mathematical treatise of SVMs can be found in Christianini (2002), Campbell (2000) and Vapnik (1995).

3.0 FEATURE SELECTION

Feature selection may be categorized into wrapper and filter models (Kohavi and John, 1991). The wrapper model selects features by directly optimizing the performance of the classifier i.e. this model seeks the subset of features yielding the best classification accuracy result. Since many subsets may have to be tried out, this approach is computationally inefficient (Kavzoglu and Mather, 2002).

Filter models are described as feature selection algorithms which use a performance metric (i.e. evaluation function) based entirely on the training data without reference to the classifier for which the features are to be selected (Kavzoglu and Mather, 2002). Examples of criterion functions encountered in remote sensing literature include statistical separability measures such as Wilk's Λ, Hotelling's T^2 and more commonly separability indices. Separability indices give a measure of how separable training data classes are hence giving an indication of how easily the dataset may be correctly classified.

Pudil and Somol (2005) further categorizes feature selection methods into optimal and suboptimal methods. Optimal methods search for the optimal subset out of all possible subsets e.g. exhaustive search methods. Suboptimal methods on the other hand make a trade off between the optimality of a selected subset and computational efficiency. Some of the search methods in this regard include greedy eliminative search, random search and guided random search methods. In this paper we consider a guided random search method called Population Based Incremental Learning (PBIL).

3.1 Exhaustive Search Method

The exhaustive search method (also called enumerative search method) works by considering all possible band combinations by way of calculating their separability indices. Although this search method guarantees the optimality of solution, it poses the problem of being computationally prohibitive (Pudil and Somol, 2005). For a dataset with *d* features (i.e. bands), $2^d - 1$ combinations are possible. This method is practicable if the number of bands is less than 10. The use of 10 or more bands would be costly in terms of computational speed. Dutra and Huber (1999) however are of the opinion that advancements in computer technology should eventually render exhaustive search an operational reality. This, including the fact that the datasets considered in this research had less than ten bands, influenced the authors decision to consider this method.

3.2 PBIL

In the random search method, a subset of features (bands) is taken at random and their separability index calculated. If many bands are being considered, the chances of hitting on the optimum subset in a limited random search will be small. However if good separability indices are possible from a variety of band combinations, there is a higher probability of encountering one of the optima quickly. The time it takes to converge to the optimum subset can be dramatically reduced by modifying the random search to a guided random search. The guided random/stochastic search method is a randomized search in which attention is adaptively increased to focus on the band combinations that return promising results. Population Based Incremental Learning is a genetic algorithm that can be used to perform a guided random search.

Genetic algorithms are motivated by the evolutionary biology principle of natural selection and genetics. From a biological context, all living organisms are made up of cells characterized by a set of chromosomes. Chromosomes consist of genes that encode peculiar traits. During reproduction, genes from parents combine to form a new chromosome with traits from either parent. According to Darwin's theory, this breeding results in the fit traits being propagated at the expense of the weaker ones which end up being filtered away. It is on this basis that genetic algorithms are tailored.

A genetic algorithm functions by randomly generating a population of 'chromosomes'. The 'chromosomes' are of equal length and may be represented by a special coding technique e.g. binary code (Tso and Mather, 2001). The fitness level of each 'chromosome' is then calculated based upon which random pairs of 'chromosomes' are bred to generate a new pool of 'chromosomes'. Breeding is effected through two mechanisms called crossover and mutation. Crossover involves the exchange of genes between two parent 'chromosomes' whereas mutation is carried out by randomly changing binary values that are representative of genes/traits. Crossover and mutation facilitate genetic diversity without which the genetic algorithm would settle into a sub-optimal state. The process of selecting and breeding define a generation. The progression of genes from one generation to another is dependent on how well the 'chromosomes' pass a fitness test. The 'chromosomes' with high fitness levels may be programmed to have a higher probability of selection to ensure that strong traits are passed on to the next generation. The number of generations may be fixed or the breeding process allowed to continue until a satisfactory level of fitness is attained.

Population Based Incremental Learning (PBIL) is an adaptation of genetic algorithms whereby the population is replaced by a probability vector instead of storing each possible solution explicitly. It is this probability vector that is updated when one progresses from one generation to another rather than the fixed population (Gosling and Cromp, 2004).

Linking feature selection to PBIL, the elements of the probability vector define the probability of selection of each feature. The idea is to use PBIL to determine the subset of bands which when classified will give as good classification results (if not better) than when all bands are utilized. In the absence of *a priori* knowledge of the importance of the bands, the probability vector is initialized to a value for example 0.5. This would mean that in a randomized selection operation, each band has an equal chance of being chosen.

In each generation a population of trials is created by sampling the probability vector with a random vector. This ensures that the inclusion of a given feature follows the probabilities in the probability vector. The fitness of each trial is determined by calculating the separability index and the trial yielding the highest index is identified as the best (chromosome in genetic algorithm technology) in that generation. Based on these results the probability vector is adjusted to reflect the best trial. If for example bands 1, 3 and 4 ended up being the best trials out of 9 bands, the probability vector corresponding to these bands would be increased slightly (by 10% from 0.5 to 0.55) while the other values would be reduced in the same proportion (from 0.5 to 0.45). This enables the subsequent generation to contain a greater proportion of trials that resemble to some degree the best trial of the previous generation.

Before proceeding to the next generation, mutation is applied to the probability vector to increase the search space in an attempt to avoid convergence towards a local optimum. Mutation may be implemented by a secondary adjustment to the probability vector in which the vector values are relaxed towards the neutral value (0.5 in this case) (Gosling et al, 2004).

Ultimately, after a series of generations the separability index of the best trial in each generation improves until the global optimum emerges. The final probability vector will also have converged towards 0 or 1 indicating the bands that had a higher or lower probability of being selected. The best trial at the end of the whole process would then represent the subset of bands with potential to yield classification accuracies comparable to those derived from using all the bands

The interest in PBIL over the traditional genetic algorithm stems from the fact that PBIL results have been found to be more accurate, and are attained faster than the traditional genetic algorithms, both in terms of the number of evaluations performed and the clock speed. This gain in clock speed is attributed to the simplicity of the algorithm (Baluja, 1994). In view of this documented simplicity and associated accuracy, PBIL was selected to facilitate feature selection n this research.

3.3 Separability Indices

The choice of adopted separability index, i.e. evaluation function, should be related to the behavior of the error made by the classifier used if the optimal subset is to be selected (Bruzzone and Serpico, 2000). Separability analysis is performed on the training data to give an indication of the expected classification error for various band combinations (Swain and Davis, 1978). However, finding an appropriate definition of interclass separability is not trivial (Schowengerdt, 1997).

In light of the fact that nonparametric classifiers are gaining prominence in remote sensing studies, there is a corresponding need for separability measures tailored around the uniqueness of these nonparametric classifiers. Whereas Kavzoglu and Mather (2002) have used the Mahalanobis distance classifier to select features for artificial neural networks (which is a nonparametric classifier like the SVMs used in this research), this research proposes a separability index tailored for the uniqueness of nonparametric classifiers. This separability measure is called Thornton's separability index.

Thornton's separability index is defined as the fraction of data points whose classification labels are the same as those of their nearest neighbors. Thus it is a measure of the degree to which inputs associated with the same output tend to cluster together (Greene, 2001). Greene (2001) expresses this concept by the following formula:

$$\text{Thornton's separability index} = \frac{\sum_{i=1}^{n} (f(x_i) + f(x_i') + 1) \bmod 2}{n} \qquad (1)$$

Where x' is the nearest neighbor of x
n is the number of points.

With this separability measure, well separated classes will have high separability indices equal to or tending towards unity. As the clusters move closer to each other and points from opposing classes begin to overlap the index begins to fall (Greene, 2001). A separability index of zero, therefore, denotes totally mixed up classes. The interest in this separability measure stems from its simplicity, speed and the fact that it is nonparametric which augers well with the SVM classifier which is also nonparametric.

4.0 STUDY AREA

To carry out this research, two study areas were considered namely: Masai Mara in Northern Tanzania and Malmesbury located in the Western Cape Province of South Africa. The Masai Mara study area was sectioned off from the Mara River Basin. Based on prior field work, six land cover classes were sought namely: wetlands, water (lakes and rivers), Bush/shrub/trees, Grasslands, "bare ground" and Roads. For the Malmesbury dataset, three land cover classes were sought for namely fields, trees and built up areas. In both cases landsat imagery was used.

5.0 METHODOLOGY

5.1 Data Processing

For each dataset, sample data corresponding to known land cover was imported into Matlab for preparation and derivation of optimum SVM parameters through cross-validation. Cross validation involves dividing the sample data into training and testing data. The training data is then used to derive kernel parameters whose measure of efficiency was determined by how well they predicted the test data classes. The kernel parameters yielding the best prediction are considered optimum and are subsequently used to classify the image. For the polynomial kernel, polynomial degree and constant C were the sought after parameters while for the radial basis function the kernel width and constant C were the required parameters which needed to be determined. The OSU_SVM Version 3.00 toolkit (developed at Oklahoma State University) was used to give Matlab SVM functionality. This toolbox was selected for its simplicity and ease of use.

5.2 Classification

SVMs are essentially a binary classification technique, however they can be adopted to handle the multiclassification tasks usually encountered in remote sensing research. To effect this, two common approaches are encountered in SVM literature. The first approach is called the 'one against one' approach and involves constructing a machine for each pair of classes resulting in N(N-1)/2 machines. When applied to a test point, each classification gives one vote to the winning class and the point is labeled with the class having most votes. The second approach involves the 'one against all' approach where by the N class dataset is divided into N two-class cases. Proponents of the ‘one against one’ approach contend that it has the advantage of avoiding highly unbalanced training data (Gualtieri and Cromp, 1998). It is however acknowledged that this approach involves constructing more SVMs as compared to the ‘one against all’ approach, hence it is more computer intensive. In this research, the ‘one-against-all approach’ was adopted. By implication therefore, different classes needed different kernel parameters. IDRISI Kilimanjaro Edition image processing software was then used to integrate the resultant class images into a land cover map. This land cover map was then georeferenced and subjected to accuracy assessment. The accuracy measure used was the overall accuracy, which gives a measure of agreement between a derived map and ground truth data.

5.3 Feature Selection

The exhaustive search method was used to compare all possible band combinations with the original band set. For both study areas 63 combinations were considered. The measure of comparison was Thornton’s separability index. For each study area and each land cover type, the separability index (S.I) of each subset was compared with the S.I of the original bands. The subset with the least number of bands yielding the closest S.I to the original bands was selected for the classification process. This subset of bands was then used to derive optimum SVM parameters mentioned previously, and subsequently used to classify the study area.

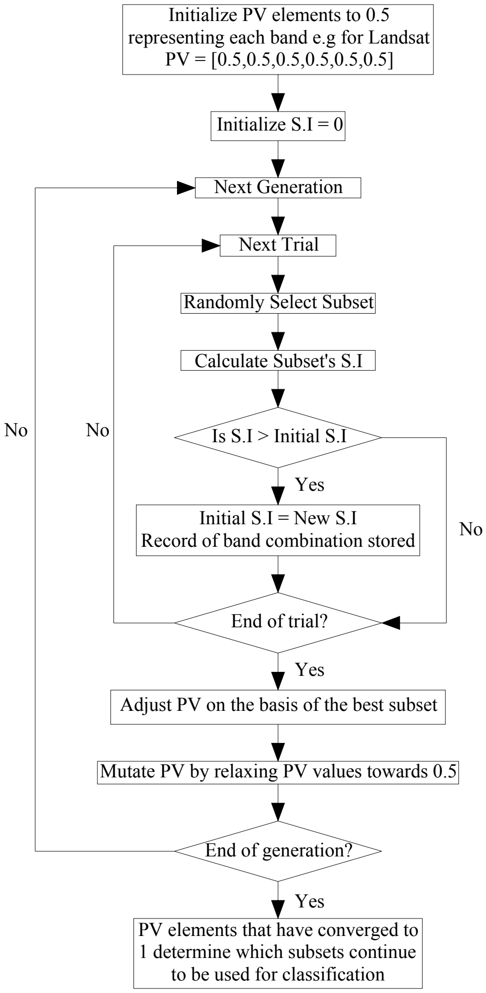
The Figure 1 shows a simplified flow chart of the algorithm used to effect PBIL classification technique.

5.4 Comparison of Feature Selection Techniques

For each study area overall accuracies were compared before and after feature selection. The significance of these differences was determined using a binomial test of significance at the 95% confidence level (critical Z value = 1.96). It is based on these comparisons that conclusions were drawn about the effectiveness and efficiency of the investigated optimization techniques.

6.0 RESULTS AND DISCUSSIONS

The Tables 1 and 2 show the selected optimum bands following the exhaustive search (ES) and PBIL for Masai Mara and Malmesbury study areas respectively. One observation from Table 1 is that ES resulted in fewer bands than PBIL. Using fewer bands for classification has the advantage of reducing the time taken to effect classification. In Table 2 on the other hand, it is only in the class ‘built up areas’ that ES has fewer bands as compared to PBIL. In the other classes i.e. ‘trees’ and ‘fields’ both ES and PBIL have resulted in identifying the same four bands.


Initialize PV elements to 0.5
representing each band e.g for Landsat
PV = [0.5,0.5,0.5,0.5,0.5,0.5]
Initialize S.I = 0
Next Generation
Next Trial
Randomly Select Subset
Calculate Subset's S.I
Is S.I > Initial S.I
Yes
Initial S.I = New S.I
Record of band combination stored
No
No
No
End of trial?
Yes
Adjust PV on the basis of the best subset
Mutate PV by relaxing PV values towards 0.5
End of generation?
Yes
PV elements that have converged to
1 determine which subsets continue
to be used for classification

PV – *Probability Vector*, SI – *Separability Index*

Figure 1: PBIL methodology

Table 1: Landsat Bands selected for Masai Mara

Class	E.S	PBIL
Wetlands	1, 2 & 6	3, 4, 5 & 7
Water	5	1 & 7
Bush/Shrub/Trees	3, 4 & 5	1, 3, 4, 5 & 7
Grasslands	3 & 4	3, 4 & 5
Bare ground	3 & 5	3, 4 & 5
Roads	2 & 5	3, 4 & 5

Table 2: Landsat Bands selected for Malmesbury

Class	E.S	PBIL
Built up areas	2, 4, 5 & 6	1, 3, 4, 5 & 6
Trees	3, 4, 5 & 6	3, 4, 5 and 6
Fields	1, 3, 4 and 5	1, 3, 4 & 5

Table 3 depicts the overall classification accuracies before (Pre-FS) and after feature selection (ES and PBIL). From this table it can be seen that for the Masai Mara dataset, overall accuracy has improved following the application of the polynomial SVM classifier to the results of ES and PBIL. On the other hand, the application of the RBF SVM classifier to both feature selection techniques has resulted in reduced overall accuracy. For the Malmesbury dataset, the polynomial SVM classifier has yielded reduced overall accuracies when applied to ES and PBIL results while the RBF SVM classifier has yielded improved overall accuracies.

Table 3: Overall accuracies for both datasets

	Masai Mara			Malmesbury		
SVM	Pre-FS	ES	PBIL	Pre-FS	ES	PBIL
Polynomial	0.67	0.71	0.75	0.69	0.65	0.66
RBF	0.83	0.64	0.79	0.65	0.66	0.66

Tables 4 and 5 give further analysis of these results by giving a summary of the significance of the differences between the overall accuracies before and after feature selection. For the Masai Mara dataset, feature selection has resulted in significantly better overall accuracies when the polynomial SVM is used, while yielding significantly worse overall accuracies when the RBF classifier is applied. In the case of the Malmesbury dataset, the differences in the overall accuracy using the results of feature selection are insignificant.

Table 4: Assessment of overall accuracies for Masai Mara

	ES		**PBIL**	
SVM	**Significance**	**Comment**	**Significance**	**Comment**
Polynomial	4.65	Significantly Better	11.07	Significantly Better
RBF	-23.74	Significantly Worse	-5	Significantly Worse

Table 5: Assessment of overall accuracies for Malmesbury

	ES		**PBIL**	
SVM	**Significance**	**Comment**	**Significance**	**Comment**
Polynomial	- 0.55	Insignificant	- 0.37	Insignificant
RBF	0.18	Insignificant	0.18	Insignificant

Of the two feature selection methods, ES is more computer intensive and hence more time consuming given that the S.Is of all possible combinations have to be evaluated before a ranking process enables the identification of the optimal subset. The efficiency of the ES technique can be improved by restricting the search to the subset of bands that is greater than the datasets intrinsic dimensionality (Dean and Hoffer, 1983). For landsat, the intrinsic dimensionality is two or three. This implies that to make ES more competitive, the search for optimal subset, only band combinations with three or more features need be considered.

On trial as well in this search was Thornton's separability index and on the balance of the results posted, it has performed well.

7.0 CONCLUSIONS

From the results, both Exhaustive Search and Population Based Incremental Learning are appropriate feature selection techniques for Support Vector Machine classification and so is Thorntorn's separability index an appropriate criterion function. The authors would like to recommend that a logical progression of this research would be to investigate appropriate feature selection techniques for emerging multi-classifier SVMs.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the German Academic Services (DAAD) and Makerere University – Uganda. They would also like to extend an arm of gratitude to the Council for Geoscience, Geological Survey South Africa who availed the images for this research.

REFRENCES

Baluja, S. 1994. ***Population Based Incremental Learning – A Method for integrating Genetic Search Based Function Optimization and Competitive Learning.*** Technical Report No. CMU-CS-94-163. (Pittsburgh, Pennsylvania: Carnigie Mellon University).

Bruzzone, L., and Serpico, S. B. 2000. A technique for feature selection in multiclass problems. ***International Journal of Remote Sensing*, 21,** 549–563.

Campbell, C. 2000. ***An Introduction to kernel Methods, Radial Basis Function Networks: Design and Applications***. (Berlin: Springer Verlag).

Christianini, N., and Shawe-Taylor, J. 2000. ***An introduction to support vector machines: and other kernel-based learning methods***. (Cambridge and New York: Cambridge University Press).

Dean, M. E., and Hoffer, R. M. 1983. Feature Selection Methodologies Using Simulated Thematic Mapper Data. In ***Proceedings of the 9th International Symposium of Machine Processing of Remotely Sensed Data***. Pp: 347 – 356, Purdue University, West Lafayette, Indiana. 21st – 23rd June 1983.

Dutra, L., and Huber, R. 1999. Feature Extraction and selection for ERS-1/2 InSAR Classification. ***International Journal of Remote Sensing*, 20,** 993 – 1016.

Foody, M. G. 2002a. Hard and soft classifications by a neural network with a non-exhaustively defined set of classes. ***International Journal of Remote Sensing*, 23,** 3853 – 3864.

Foody, M. G. 2002b. Status of land cover classification accuracy assessment. ***Remote Sensing of Environment*, 80,** 185 – 201.

Gilad-Bachrach, R., Navot, A., and Tishby, N. 2004. Margin based Feature Selection – Theory and Algorithms. In ***Proceedings of the 21st International Conference on Machine Learning,*** Banff, Alberta, Canada. 4th – 8th July 2004.

Gosling, T., Nanlin, J., and Tsang, E. 2004. ***Population Based Incremental Learning Versus Genetic Algorithms: Iterated Prisoners Dilemma***. Technical Report No. CSM-401. (Essex: University of Essex, England).

Greene, J. 2001. Feature subset selection using Thornton's separability index and its applicability to a number of sparse proximity-based classifiers. In ***Proceedings of the 12th Annual Symposium of the Pattern Recognition Association of South Africa.*** November 2001

Huang, C., Davis, L. S., and Townshed, J. R. G. 2002. An assessment of support vector machines for land cover classification. ***International Journal of Remote Sensing,* 23,** 725–749.

Hughes, G. F. 1968. On the mean accuracy of statistical pattern recognizers. ***IEEE Transactions on Information Theory,* 14,** 55–63.

Joachims, T. 1998. Text categorization with support vector machines—learning with many relevant features. In ***Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany.*** (Berlin: Springer), pp. 137–142.

Kavzoglu, T. and Mather, P. M. 2002. The role of feature selection in artificial neural network applications. ***International Journal of Remote Sensing,* 15,** 2919–2937.

Kohavi, R., and John, G. 1997. Wrapper for feature subset selection. ***Artificial Intelligence,* 97,** 273 – 324.

Mahesh P., and Mather, P. M. 2003. Support Vector classifiers for Land Cover Classification. In ***Proceedings of the 6th Annual International Conference, Map India 2003, New*** Delhi, India. 28th – 31st January 2003.

Mashao D. 2004. Comparing SVM and GMM on parametric feature-sets. In ***Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa***, Cape Town, South Africa. 27th – 29th November 2004.

Muasher, M. J., and Landgrebe, D. A. 1982. A binary tree feature selection technique for limited training sample size. In ***Proceedings of the 8th International Symposium of Machine Processing of Remotely Sensed Data***. Purdue University, West Lafayette, Indiana. 7th – 9th July 1982. Pp. 130 – 137

Pudil, P., and Somol, P. 2005. An Overview of Feature Selection Techniques in Statistical Pattern Recognition. In ***Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*** Nov 2005.

Schowengerdt, R. 1997. ***Remote Sensing: Models and Methods for Image Analysis*** (2nd Edition). (San Diego: Academic Press).

Swain, and P.H., and Davis, S. M. 1978. ***Remote Sensing: The Quantitative Approach.*** (New York : McGraw-Hill).

Tso, B., and Mather, P. 2001. ***Classification Methods for Remotely Sensed Data***. (London and New York: Taylor and Francis).

Vapnik, V. N. 1995. ***The Nature of Statistical Learning Theory***. (New York: Springer-Verlag).

Wu, D., and Linders, J. 2000. Comparison of three different methods to select features for discriminating forest cover types using SAR imagery. ***International Journal of Remote Sensing,* 21,** 2089 – 2099.