Classification of RS data using Decision Tree Approach

Pooja A P
M Tech student, Dept of ECE
Malnad College of Engineering, Hassan, Karnataka, India

Jayanth J
Lecturer, Dept of ECE
G.S.S.S.I.E.T.W, Mysore, Karnataka, India

Dr Shivaprakash Koliwad
Professor and Head (R), Dept of ECE
Malnad College of Engineering, Hassan, Karnataka, India

ABSTRACT
The traditional hard classification techniques are parametric in nature and they expect data to follow a Gaussian distribution, they have been found to be performing poorly on high resolution satellite images, as classes in these images tend to exhibit extensive overlapping in spectral space. This produces spectral confusion among the classes and results in inaccurate classified images. A major drawback of such classifiers lies in the difficulty of integrating ancillary data, which follows a non Gaussian distribution nature. Ancillary data provides extra spectral and spatial knowledge, which improves the classification accuracy. Classification done using such knowledge is known as knowledge base classification. The present study explores a non-parametric decision tree classifier to extract knowledge from the spatial data in the form of classification rules. The classified image overall accuracy was found to be 86.66% using the Decision Tree method and with kappa values .8133 respectively.

Keywords
Decision Tree Classifier (DTC), Remote Sensing (RS), Maximum Likelihood Classifier (MLC), Image Classification

1. INTRODUCTION
Classification of a remotely sensed (RS) image can be seen as an iterative process in which each of its pixels is assigned to one of the several predefined land cover classes to be mapped. The goal of image classification is to exploit the spectral, spatial and temporal resolution of data and other characteristics such as multi polarization, multi frequency and multi incident angle signature to make the classification more reliable and accurate. The Image classification is a process used to produce thematic maps for digital imagery [5]RS data classification is based on a unique relationship between a given materials or land cover class and its reflected radiation at certain wavelength (reflectance) contained in a spectral band of an image, i.e., a one-pixel–one-class relationship [2]. There are two approaches namely unsupervised classification and supervised classification. They are also known as hard classifiers.

These classifiers are parametric in nature, and they examine only the spectral variance ignoring the spatial distribution of the pixels in the satellite data. They perform poorly on high resolution satellite images, due to mixed pixel problem. This is due to the fact that the pixel resolution fails to correspond to the spatial characteristics of the target. But if the spatial resolution of the satellite imagery is increased, the spectral values of the pixels tend to overlap [3]. Hard classifiers give reliable results if the training sites, as well as the image, are both very homogeneous. They tend to over generalize the resulting land cover map if the number of clusters is small. They may fail because the spectral signature of a given land cover may be too general to describe properly all the pixels considered to be a part of it. As spectral signature is a statistical description of the reflectance of a land cover type in every spectral band considered [4]. Thus they have limited success on high resolution multispectral images.

To overcome this shortcoming knowledge based classification technique is used. They are non parametric in nature and are not dependent on the way in which the data is distributed; hence it is suitable for incorporation of non-spectral values into classification procedure [1]. In this method attributes such as region shape, size, texture information, and elevation are included; along with it human expert's knowledge is organized in a knowledge base to be used as an input of automated interpretation processes. This automated interpretation process enhances the classification accuracy and performance. Therefore, knowledge-based image interpretation systems arise as an effective tool for image interpretation [8]. Some of the methodologies are Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree Classification (DTC). ANN and SVM exhibit higher accuracy, but they are relatively slow compared to that of Decision Tree Classifier (DTC). The training period increases as the size of the training data is increased [7] [9]. They have also concluded that training a decision tree classifier is much faster and are easy to analyze, than compared to other classifiers [8]. In literatures it is found that Decision trees perform better than other classification algorithms [1] [10] [3]

In our work decision tree approach is adopted to classify a multispectral satellite image. The C4.5 classification algorithm proposed by J Quinlin is been used [11]. Accuracy assessment is performed over the classified image, an overall classification accuracy of 86.66% and kappa statistics of .8133 is obtained. For evaluating performance, the satellite data classified using Maximum Likelihood Classification (MLC). An improvement in accuracy is seen using DTC.

2. DECISION TREE CLASSIFIER
Decision tree (DT) is one of the inductive learning algorithms that generates classification tree using the training data/samples. It is based on the "divide and conquer" strategy [6]. It is a non parametric in nature hence independent of the properties of the distribution of data, thus suitable for incorporation of non-spectral data into classification procedure so improvement in class separability can be achieved .The resulting decision tree provides a representation of the concept that appeal to human

because it renders the classification process self-evident. It supports classification problems with more than two classes and can be modified to handle regression problems [6] [12].

DT follows a hierarchical structure where at each level a test is applied to one or more attribute values that may have one of two outcomes. In order to classify an object, we start at the root of the tree, evaluate the test, and take the branch appropriate to the outcome. The process continues until a leaf is encountered, at which time the object is asserted whether it belongs to the class named by the leaf. Each final leaf will be the result of following set of mutually exclusive decision rules down the tree. The tree is expanded until every training instance is correctly classified; over fitting of data is avoided by pruning the training dataset.

Decision trees are sometimes more interpretable than other classifiers such as neural networks and support vector machines because they combine simple questions about the data in an understandable way [6]. Decision tree approach has substantial advantages for land use classification problems because of there flexibility and ability to handle non-linear relations between features and classes, hence improves the classification accuracy to a great extent [1]. The major drawback of DTC technique is they are unstable when feature space and training areas are changed. A sample decision tree model is shown in the figure1.

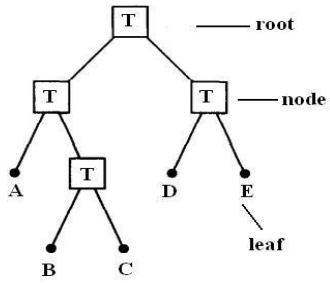

Figure1: Decision Tree

The generalized method for constructing a decision tree can be summarized as follows:

- If there are *k* classes denoted $\{C_1, C_2,..., C_k\}$, and a training set, *T*, then
- If *T* contains one or more objects which all belong to a single class *Cj*, then the decision tree is a leaf identifying class *Cj*.

If *T* contains no objects, the decision tree is a leaf determined from information other than *T*. If *T* contains objects that belong to a mixture of classes, then a test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes $\{O_1, O_2,..., O_n\}$. *T* is partitioned into subsets $T_1, T_2,..., T_n$, where *Ti* contains all the objects in *T* that have outcome *Oi* of the chosen test. The same method is applied recursively to each subset of training objects to build the decision

3. C4.5 CLASSIFICATION ALGORITHM

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [11]. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. Information gain ratio is used as the splitting criteria in this algorithm. The splitting ceases when the number of instances to be split is below a certain threshold [14]. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from selecting an attribute for splitting the data. The attribute with the highest normalized information gain is selected to make the decision. The C4.5 algorithm then recurs on the smaller sublists. Let S be a given training set then,

$$Informationgain(a_i, s) = Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)_i} \frac{\left|\sigma_{a_i = v_{i,j}} S\right|}{|S|} \bullet Entropy(y, \sigma_{a_i = v_{i,j}} S)$$

$$Entropy = (y, S) = \sum_{c_j \in dom(y)} -\frac{\left|\sigma_{y = c_j} S\right|}{|S|} \log_2 \frac{\left|\sigma_{y = c_j} S\right|}{|S|}$$

$$Gainratio(a_i, S) = \frac{Informationgain(a_i, S)}{Entropy(a_i, S)}$$

The thresholds used for each nodal decision are selected using minimum entropy or minimum error measures. It is based on using the minimum number of bits to describe each decision at a node in the tree based on the frequency of each class at the node.

The pruning method implemented in this algorithm is Error Based Pruning (EBP) [13] [14]. Information in the training set is used for building and simplifying trees. Here pruning is performed based on bottom-up post-order traversal strategy. The specialty of EBP is that it simplifies the decision tree by grafting a branch onto the place of the parent root, along with pruning of nodes [14]. The C4.5 algorithm incorporates missing values in the training data by using corrected gain ratio criteria [13]. J4.8 algorithm is Java implementation of C4.5 algorithm. This classifier built on J4.8 algorithm is called the J4.8 DTC. It incorporates information gain ratio as attribute splitting criterion [15]. This algorithm is used in our present work for learning and induction of classification rules from pruned partial decision trees built using C4.5. The detailed description of these algorithms is available in [15] [11].

4. STUDY AREA

The satellite data used for the present study is LISS III sensor image of IRS-P6. It is a multispectral data having band resolution of Green: (0.52-0.59μm); Red: (0.62-0.68μm); Infrared: (0.77-0.86μm) and a spatial resolution of 5.8m. The study area considered for the present work is a segment of semi urban area of Kumta town. The region of interest Kumta is located on the Arabian sea coast in the district of Uttara Kannada in the state of Karnataka. Kumta is geographically located at 14°25'North and 74°24'East. It has an average elevation of 3 meters from sea level. Kumta Taluk is a semi urban area, it has a good mixture of spectrally overlapping classes comprising of man made structures and natural land cover features.

5. METHODOLOGY

The raw satellite data is subjected to image pre-processing. The satellite data is geo referenced to real world co-ordinates; this is done in order to align the satellite image with that of the actual co-ordinates of the real world. This is known as geometric correction [2].

The training data is prepared using the spectral (RGB) values of pixels of the satellite image. Training data set consisting of 2178 instances was prepared. Weka data mining tool is used for generation of decision tree from which the classification rules are derived. J4.8 algorithm is used in our work for tree generation. Ten fold cross validation test mode is selected for training the dataset.

The rules generated are implemented on the satellite image data for classification. This work was carried out in Erdas Imagine V 8.5 RS image processing software. A knowledge base is created using the classification rules, DTC classification was performed using the rules. The satellite data is also classified using MLC using the same training set. Totally eleven classes were defined, and their signature file was created. MLC classification was carried out over the satellite image using the signature file. Accuracy assessment was carried out with a validation dataset size of 256 instances for both the methods. The obtained results were compared to evaluate the performance of DTC with that of MLC.

6. ANALYSIS AND RESULTS

In this section we discuss the results obtained after carrying out the classification procedure on the selected study area. A comparison of the performance of the DT algorithm with that of the MLC has been stated here.

The training sample consisting of 2178 instances was fed into the J4.8 decision tree algorithm to generate classification rules. A total of 48 rules were generated. Some of the rules generated are:

- If green <=47 and red <=68 and red<=18 and blue <=83 and red <=12 → SEA WATER
- If green > 47 and blue <= 107 and red <= 92 and red <= 54 and blue <=86 and red <=40 →SUBMERGED AREA
- If green <=47 and red <= 68 and red >18 and blue<=79 and red <=43 and blue <=69 →STONE
- If green > 47 and blue <= 107 and red <= 92 and red > 54 and blue > 103 and green <= 77 →HOUSE
- If green <= 47 and red <= 68 and red > 18 and blue <= 79 and red > 43 and red > 60 and blue > 66 and green > 42 → GRASS DRY_AREA
- If green > 47 and blue <= 107 and red > 92 and green <= 62 and red > 128→GRASS LAND
- If green<=47 and red<=68 and red>18 and blue <=79 and red <=43 and blue >69 and green <=35→POOL
- If green<=47 and red<=68 and red>18 and blue <=79 and red >43 and red >60 and blue <= 66→TREES
- If green > 47 and blue <= 107 and red <= 92 and red > 54 and blue >103 and green <= 77 →PLAIN LAND
- If green <= 47 and red <= 68 and red<=18 and blue <=83 and red > 12 → RIVER

A knowledge base was created using these classification rules. This knowledge base is extracted over the satellite image to obtain the decision tree classified image. The legend/ signature file used for classification is shown in Table1. The Figure3 is the classified image obtained by applying DTC method. The satellite image was also classified using MLC. The classified image obtained by applying MLC is shown in Figure4.

Figure 2: Study Area [Multispectral Image (resolution=5.8m)]

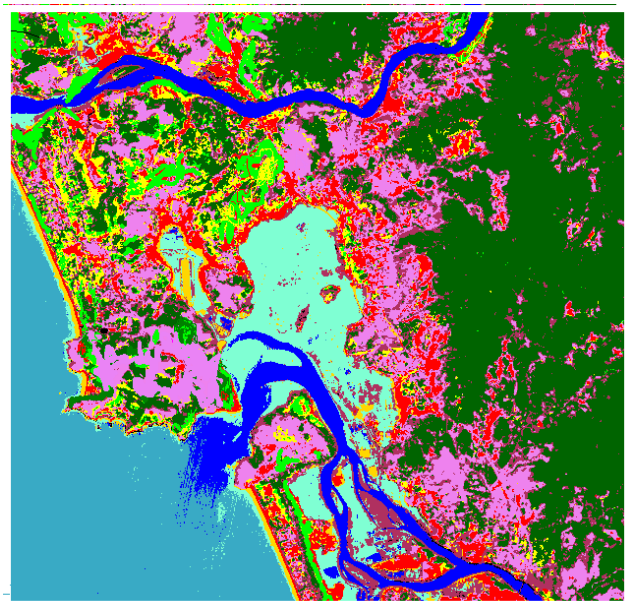

Figure 3: DTC Classified Image

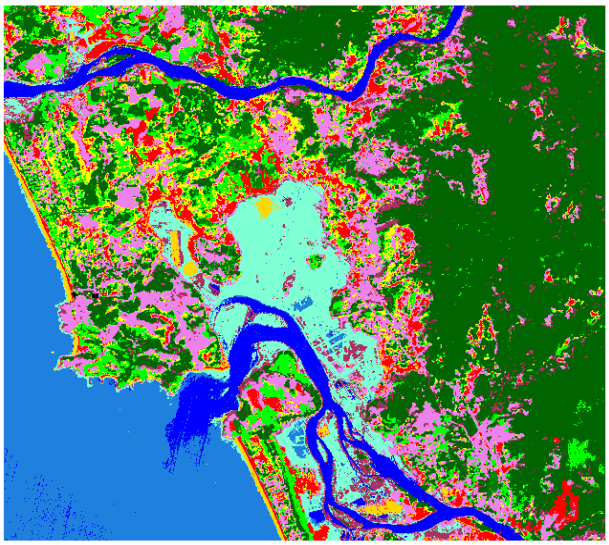

Figure 4: MLC Classified Image

Table 1. Legend/ Signature file used for classification

Class name	Colors used to represent
Stone	
House	
Grassland	
Grass dry area	
Plain land	
Sand	
River	
Submerged area	
Sea water	
trees	
pool	

Accuracy assessment was carried out over both the classified image, the overall classification accuracy and kappa statistics were evaluated for both the cases. The OCA obtained by decision tree method was 86.66% and that with MLC was 81.96%. The kappa statistics obtained for DTC is 0.8133 and that for MLC is 0.7653. The accuracy details are listed in the following tables.

From the results it can be noticed that there is misclassification seen in the class stones and house, none of the classifiers have performed better here. Misclassification is noticed between the classes' stones and trees this is due to the highly overlapping spectral values due to the shadow of trees, misclassification is also observed between pool and river, again due to overlapping of spectral values. From the classified images and conditional kappa values its clear that few of the classes like sand and grassland have shown very good results in DTC. The classes which have uniform distribution over the study area have been classified correctly in both the classifier, than those classes which are scattered in the study area. For these scattered classes DTC performance is better than that of MLC.

DTC is easy to analyze than MLC as the classification rules generated are simple to understand and implement. Also the computation time required for training the test sample is less than the other approach. In an overall analysis almost all the classes perform better under DTC than that of MLC. An improvement of 4.8% in the classification accuracy is obtained in DTC technique.

Table 2. Comparison of Classification Accuracy

	Overall classification accuracy	Overall kappa statistics
DTC	86.66%	0.8133
MLC	81.96%	0.7653

Table3. Accuracy Table for MLC

Class name	Reference total	Classified total	Number correct	Producer's accuracy	User's accuracy
Stone	4	8	3	75.00%	37.50%
House	11	17	9	81.81%	52.94%
Grassland	11	7	6	54.54%	85.71%
Grass dry area	15	11	11	73.33%	100%
Plain land	19	18	14	73.68%	77.78%
Sand	3	2	2	50.00%	100%
River	15	15	14	93.33%	93.33%
Submerged area	17	28	15	88.24%	83.33%
Sea water	39	38	37	94.87%	97.36%
Trees	95	93	89	93.68%	95.69%
Pool	4	2	2	50.00%	50.00%

Table4. Accuracy Table for DTC

Class name	Reference total	Classified total	Number correct	Producer's accuracy	User's accuracy
Stone	11	26	10	90.90%	38.46%
House	5	10	4	80.00%	40.00%
Grassland	11	7	6	54.54%	85.71%
Grass dry area	36	37	29	80.55%	78.37%
Plain land	13	15	11	84.62%	73.33%
Sand	5	3	2	33.33%	66.66%
River	19	13	11	57.89%	84.62%
Submerged area	20	22	19	95.00%	86.36%
Sea water	43	44	43	100%	97.72%
Trees	88	76	73	82.95%	96.05%
Pool	4	2	1	25.00%	50.00%

Table5. Conditional Kappa for each class

	DTC	MLC
Stones	0.4356	0.3596
Trees	0.8951	0.9396
Grassland	0.8677	0.5419
Grass Dry area	0.6959	0.7550
Submerged area	0.8214	0.8521
Sand	1.0000	0.4941
Seawater	1.0000	0.9472
River	0.9292	0.8338
Plain Land	0.7599	0.7191
House	04817	0.3880
Pool	0.5152	0.4513

7. CONCLUSION

In our paper, decision tree classification approach for remotely sensed satellite data is developed and implemented. The reason for high accuracy may be to some extent attributed for the reason that the part of the training set is being considered as ground truths instead of actual data. DTC can be trained quickly and it also acquires less computational time. It can be concluded from the study that decision tree classification algorithm perform better than MLC as DTC does not depend on a prior model, it is dynamic in nature. But the problem associated with decision tree is they are unstable when the feature space or the training data is changed. Larger the training set, greater is the accuracy achieved. The accuracy of the results depends only upon the test set selected; the efficiency of any algorithm should not be decided on the accuracy measure alone.

8. ACKNOWLEDGEMENTS

It is a pleasure to recognize the many individual who have helped me in completing this technical paper. I sincerely express heartiest thanks to Dr A S Ravikumar (Associate Professor, dept of Civil, UVCE Bangalore) and Dr. Ashok Kumar (Principal V.I.T Puttur) for all the technical guidance, encouragement and analysis of the data throughout this process.

9. REFERENCES

[1] Mahesh Pal & Paul M Mather, "Decision Tree Based Classification of Remotely Sensed Data", *centre for remote imaging, sensing and processing (CRISP)*, November 2001.

[2] Paul M Mather, "Computer Processing of Remotely-Sensed Images- An Introduction", Wiley Publications, 3rd edition, 2004.

[3] D.LU & Q.Weng, "A Survey of Image Classification Methods and Techniques for Improving Classification Performance", International Journal of Remote Sensing, Vol.28, No.5, 10 March 2007, 823-870.

[4] Andras Bardosy & Luis Samaniego "Fuzzy Rule-Based Classification of Remotely Sensed Imagery", IEEE Transactions on Geoscience and Remote Sensing, Vol.40, No. 2, February 2002.

[5] Robert A Schowngerdt, "Remote Sensing Models & Methods for Image Processing", 2nd Edition, Elsevier Publications, 2006.

[6] Rasoul Safavian & David Landgrebe, "A survey of Decision Tree Classifier Methodology", IEEE Transactions on Systems, Man and Cybernetics, Vol.21, No.3, May/June 1991, pp.660-674.

[7] Mahesh Pal & Paul M Mather, "A comparison of Decision Tree and Back Propagation Neural Network Classifiers for Land Use Classification", Geo science and remote sensing symposium, 2002, IEEE International proceedings, vol. 1, pp. 503-505.

[8] C. Apte & S. Weiss, "Data Mining with Decision Trees and Decision Rules, future generation computer systems", November 1997, pp. 197-210.

[9] J R Otukei & T Blaschke, "Land Cover Change Assessment Using Decision Trees, support vector machines and maximum likelihood classification algorithms A survey of image classification methods and techniques for improving classification performance", International Journal of Applied Earth Observation and Geo information 12S, pp.S27-S31, 2010.

[10] Brandt Tso & Paul M Mather, "Classification Methods for Remotely Sensed Data", 2nd edition, CRC Press 2009.

[11] J R Quinlan, "Decision trees and decision making", IEEE transactions on Systems, Man and Cybernetics, Vol.20, No.2, March/April 1990, pp.339-346.

[12] Sam Drazin & Matt Montag, Decision Tree analysis using WEKA, university of Miami.

[13] Lior Rokach & Oded Maimon, "Top Down Induction of Decision Trees Classifiers- A Survey", IEEE Transactions on systems, Man and Cybernetics-Part C: Applications and Reviews, vol 35, No. 4, pp. 476-487, November 2005.

[14] Florina Esposito, Donato Malerba & Giovanni Semeraro, "A Comparitive Analysis of Methods for Pruning Decision Trees", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5 pp. 476-491, May 1997.

[15] H. Ian Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", Elsevier, 2nd edition,2005