DATA MINING TECHNIQUES FOR LAND USE LAND COVER CLASSIFICATION USING MULTI-TEMPORAL AWIFS DATA

Sreenivas Kandrika* and Roy, P.S.#

RS&GIS Applications Area
National Remote Sensing Agency
Balanagar, Hyderabad-500037, India
* sreenivas_k@nrsa.gov.in, kandrika@gmail.com # psr@nrsa.gov.in

Commission VII, WG VII/5

KEY WORDS: Land Cover, Land Use, Classification, Data mining, AWiFS, Multitemporal.

ABSTRACT:

The present study addresses the attempt made to explore the temporal (5-day revisit) and spatial resolution (56m) potential of AWiFS sensor aboard IRS-P6 to generate the land use land cover information using decision tree classification technique using See 5 data mining algorithm. The results obtained after two annual cycles and issues related to digital classification of temporal satellite data were presented and discussed. The temporal datasets were co-registered to sub-pixel accuracy and were atmospherically corrected using modified dark pixel method. Scaled reflectance values were extracted for various classes and rule sets were generated using See-5 data mining algorithm. These rule sets were ported into ERDAS Imagine Knowledge Engineer and the temporal data sets were classified. The results indicate that temporal satellite data at monthly interval found to be suitable to address the seasonal variability in agricultural cropland. The problem with temporal dynamics of cloud cover could be overcome with a little extra care during training site selection. Additional training sites should be defined in cloudy regions keeping its temporal dynamics of the target class in view. Mis-registration among temporal data sets too can influence classification accuracies. Among various land cover classes, classification accuracy is poorer in classes those devoid of vegetal cover. Overall kappa statistic was 0.866 for 2004-05 which was further improved to 0.908 during 2005-06.

1. INTRODUCTION

Land use land cover (LULC) information is one of the most important spatial information that is often required and got an important role in natural resource management and environmental planning. Till recent past, limited satellite data sets were used to derive land cover information. This approach often suffered with its inability to explain the staggered sowing across cropping seasons especially in agricultural areas. Multi-dimensional data sets were used to derive the land cover information at various scales.

Globally there are several land use land cover classification systems in vogue. These systems were developed to address specific project goal. In India, since major cropping patterns are more or less stabilized. Thus first cut information on net sown area will be very useful for taking stock of country's agricultural situation. Keeping this in view, the land cover classification addressed in the present study incorporates agricultural classes discernable in a regional scale mapping programme. However, the agricultural cropping pattern will be very much staggered overlapping major cropping seasons namely, *kharif* (June to November), *rabi* (November to March) and *zaid* (March to June). Added to this, the duration of various crops ranges from 40 days to 120 days besides long duration crops (more than 6 months) like sugarcane, pigeon pea and cotton. Thus to address such a complex cropping pattern with a cloud free data sets, the satellite data need to collected at a shorter intervals.

When it comes to deriving information from satellite data through digital classification approach, supervised classification employing maximum likelihood algorithm has been the most commonly used digital classification technique on remotely sensed data (Richards, 1993). This classification method assumes that the probability distributions for the input classes are in multivariate normal form. This poses a limitation when spatial data with non-Gaussian distribution is also included while describing the data dimensionality. A common problem while using prior probabilities especially with maximum likelihood algorithm is that they can bias the posterior probability of a class towards the result predicted by the ancillary information (Strahler, 1980). Hence, nonparametric classification algorithms are being increasingly used, which make no assumptions regarding the distribution of the data being classified (Foody, 1997; Carpenter et al., 1999). Nonparametric classification techniques found to be useful for land cover mapping when there is substantial intra-class variability and when the land cover classes tend to be multimodal (Gopal, et al, 1999; Friedl et al., 2000; Hansen et al, 2000).

The computational simplicity and operational flexibility of nonparametric methods facilitated land cover classification significantly. One such widely used nonparametric classification technique is decision trees. Several authors demonstrated the utility of decision trees derived in supervised fashion provide an accurate and efficient way for land cover classification problems using remote sensing data (Friedl and Brodley, 1997; Hansen et al., 1996; Swain and Hauska, 1977). The decision trees can be used to handle the nonlinear relationships among input data sets (Xu et al, 2005) and noisy as well as missing data (Quilan, 1993).

Keeping the land use land cover information requirement in view, the present article addresses the suitability of commercially available See-5 decision tree (DT) classifier to handle the temporal spectral variability to capture land use and land cover information. The results obtained after two annual cycles of multidate analysis and issues related to digital classification of temporal satellite data using decision trees were presented and discussed. In this study an effort was made to addresses the impact of sample size on decision tree and its accuracy, effect of satellite data mis-registration and utility of NDVI as a stand alone parameter for land use land cover classification.

2. METHODOLOGY

The entire process adopted in the present work involves top of atmospheric reflectance generation, atmospheric correction, georectification, data mosaicing, ground sample collection, generation training sets from multi-temporal AWiFS data and digital classification using rules generated from See-5 DT classifier and accuracy assessment.

2.1 Pre-processing

The digital numbers (DNs) of all the data sets were converted into at-satellite radiance values using the linear scaling functions provided in the data product's leader file. Using the orbit parameters like date of acquisition, solar zenith angle and Earth-to-Sun distance, top of atmospheric reflectance has been computed as per Liang *et al* (2001). The solar irradiance values for the AWiFS spectral bands were computed using the solar tables provided by Nickel and Labs. Finally, these outputs were re-scaled for 10-bit output using a constant scaling function across bands. The temporal data sets were georeferenced using image to image tie down procedure. The georeferenced master reference image already available has been used for georectification though image-to-image tie down procedure. For all the data sets used in the present study, the individual RMS error of each data set was < 50m and was in range of 26m to 50m. The atmospheric correction was carried out using modified dark pixel subtraction method (Chavez, 1988).

2.2 Ground truthing

Historic as well as current satellite data has been used along with legacy maps for planning the ground truth campaign. Historic satellite data helped in locating the hotspots where land cover changes have taken place. During *kharif* as well as *rabi* crop seasons, ground truth was collected covering various land cover types with a special focus on agricultural crops. To record the location of various ground truth sites, EMTAC Bluetooth GPS was used. A part of these points were used for defining the training areas, while a portion was used for accuracy assessment.

2.3 Digital Classification

Besides using the scaled reflectance (SR) values per se, an attempt was also made to study Normalized Difference Vegetation Index (NDVI) as a standalone parameter for LULC classification. NDVI images were generated using simple image arithmetic.

$$NDVI = (NIR - Red) / (NIR + Red) \quad (1)$$

Where,

NIR and Red are radiometrically normalized and atmospherically corrected reflectance values of NIR and Red bands of AWiFS.

Initially the temporal spectral response of various land use land cover classes was studied. Training sets were defined in light of their temporal response vis-à-vis ground truth information by contemporaneous visual inspection of temporal images. Digital counts (scaled reflectance values) as well as NDVI values were extracted for these training areas defined for various classes. Rules set / decision tree was generated using See-5 data mining software for various sample sizes.

The resultant accuracies of training set classification were studied along with the rule sets derived. These decision trees were analyzed in light of their characteristics and accuracies. These rule sets were ported into ERDAS Imagine Knowledge Engineer and the temporal data sets were classified. To accomplish the workflow, an interface developed with Visual Basic 6.0 was used.

3. RESULTS & DISCUSSION

The results obtained from the present study were arranged into three sub-sections viz., effect of sample size, utility of NDVI as a stand alone parameter, effect of mis-registration on land use land cover classification.

3.1 Temporal signatures

The temporal spectral response of various target classes indicate that, vegetation by virtue of having low reflectance in red and high reflectance in NIR forms a characteristic curve which clearly separates from other land cover classes like bare fields and water bodies. The temporal combination of the above mentioned spectral reflectance patterns distinctly separates various land cover classes. This especially helped in separating various seasons of cropping – *kharif*, *rabi* and *zaid*. While *kharif* crop exhibits vegetation reflectance pattern during October, and exhibits a bare soil reflectance curve during subsequent periods. While *rabi* crop portrays characteristic vegetation signature during November to March, it shows characteristics response of bare soils during other months. The triple cropped areas, agricultural and forest plantations and forest areas do have an overlapping temporal signature with distinct spectral response pattern in at least one of the data sets which helped their separation. However, overlapping temporal and spectral response between plantation and semi-evergreen forest present this area resulted in poor classification accuracies. This could be efficiently overcome by including a mask separating forest and non-forest areas in the classification.

3.2 Effect of training sample size

An attempt was made to understand the influence of training set size on classification accuracy and the number of rule sets generated by decision tree classifier. Initially large training sets (approximately 2500 pixels) covering various possible variations in temporal response were defined for each class and the SR values were extracted. The variation in training set definition by virtue of cloud cover was treated as a separate class. From this pool of training set pixels, training samples of

various land cover classes were drawn randomly in the multiples of 50. The influence of training set size was tested up to 750 pixels per class. It was observed that there was an improvement in training set classification accuracy up to 300 pixels per training set and there after it is more or less stabilized in cloud free areas.

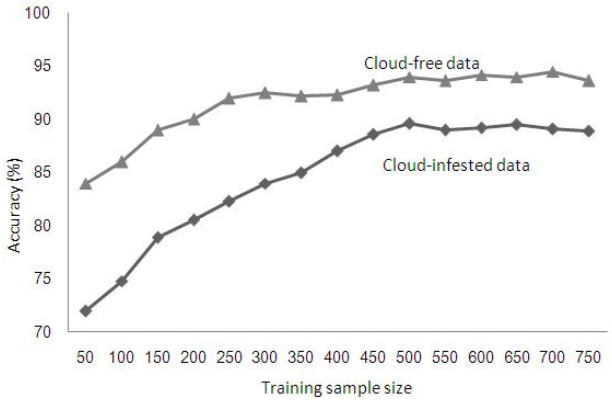

Figure-1. Effect of training sample size on classification accuracy of evergreen forest class.

The cross computed classification accuracy of training sets was found to increase as the number of training set size increases. The general range of classification accuracy is 74.8% to 97.3%. In figure-1 the effect of training sample size for a representative training class like evergreen forest has been depicted. For most of the classes, this increasing trend was observed up to a training set size of 300 pixels and there after showing an insignificant change in accuracy. However for cloud infested areas, the training set accuracies were improving up to 500 pixels size. Thereafter there were minor deviations in accuracy which may be due to intra-class variability of training set or may be because of random selection of training pixels from pool. This indicates that the training set size has got influence on accuracy and up to a certain sample size. This sample size however can not be fixed as same for all the land cover classes. It depends on the inherent variability in the spectral response of target features. To address the missing data (like cloud infested areas), the sample size need to be increased.

3.3 Influence of cloud contamination

The cloud contamination in the data sets has influence on the decision tree structure (table-1).

Sl. No	Description	Max. Tree depth	No. of nodes	No. of leaf nodes
1	Cloud free data	4	116	82
2	Cloud infested areas	6	162	101

Table-1. Decision tree structure as a function of cloud contamination.

From the above table it is clear that the cloud contamination can influence the decision tree structure by increasing the tree depth as well as number of nodes. From the above table it could be observed that the number of nodes used for branching the decision tree (total nodes – leaf nodes) as well as tree depth is high in cloud infested areas. The same has also been observed in the targets that have heterogeneous target representation by means of temporal staggering in spectral response.

3.4 NDVI as a stand alone parameter

The decision tree classification has been performed on temporal data sets with SR as well as NDVI values alone and in combination (SR + NDVI together). There was no significant improvement in overall classification accuracy when NDVI is added to SR values over using SR values per se. Using temporal NDVI values as stand alone data set resulted in significantly poorer classification accuracies. Thus NDVI as a stand alone parameter is not very much suitable for LULC digital classification using DT approach.

3.5 Effect of mis-registration on land use land cover classification

The co-registered temporal datasets do have mis-registration across data sets ranging from 1 to 3 pixels. This resulted in mis-classification in the output especially in borders where the transition between land cover classes happens. However, the misclassification is not the same across spatial domain. Further, the extent of mis-classification depends on the decision rule used while classifying a pixel.

Accuracy assessment

The classified output was subjected post classification accuracy assessment. A stratified random sampling approach has been adopted with at least 15 ground reference points per class. The location of each point has been recorded with the help of GPS. The output class for each of these points has been extracted from classified image and a confusion matrix has been generated. The following table provides the conditional kappa statistics of various LULC classes.

Sl. No	LULC class	2004-05	2005-06
1	Built up land	0.9685	0.9715
2	Kharif crop land	0.8883	0.8937
3	Rabi crop land	0.8653	0.9467
4	Zaid crop land	0.9015	0.9467
5	Double / triple crop land	0.8480	0.8933
6	Current fallow land	0.8141	0.8400
7	Plantations / orchards	0.8998	0.8400
8	Evergreen / Semi-Evergreen forest	0.9638	0.9465
9	Deciduous forest	0.7908	0.8416
10	Shrub or degraded forest	0.8293	0.9685
11	Swamp / Mangrove	0.9060	0.9467
12	Grassland & Grazing Land	0.8587	0.9918
13	Other Wastelands	0.8913	0.8405
14	Gullied/Ravines	0.8295	0.9461
15	Scrubland	0.8651	0.8416
16	Water bodies	0.9481	0.8416
17	River sand	0.9815	0.9981

Table-2. Class-wise conditional Kappa statistics.

The lower classification accuracy in deciduous forest is due to the signature overlap with shrub / degraded forest class. Similarly, there exists a signature overlap among other wastelands class, current fallow and gullied / ravenous areas. Overall kappa statistic for the classified output was 0.866 during the year 2004-05. It was further improved to 0.908

during 2005-06 due to availability of relatively cloud-free data set covering most of the variation in staggering of crops.

4. CONCLUSIONS

The decision tree classification algorithm (See-5) used in this study is able to exploit the temporal variation in target spectral properties satisfactorily. The decision tree classification could effectively tackle the temporal variability and the output image is having relatively low salt-pepper noise and is spatially contiguous.

There is high temporal variation of crop cover especially across dryland and irrigated regions by virtue of variation in crop type and staggered sowings. Results indicate that temporal satellite data at monthly interval found to be suitable to address the variation cropland. The temporal data was found to be very useful in separating fallows with sparse grass cover during *kharif* season and crops those are in initial vegetative phase. It was also found that the problem with temporal dynamics of cloud cover could be overcome with a little extra care during training site selection. However, there was a relatively lower accuracy in permanent vegetation classes when compared to areas those with temporally variable vegetal cover. These classes include evergreen forests, and very dense scrubs with broad-leaved vegetation in northern hilly region and plantations in plains. The plantations could be separated from forests by addition of forest mask into the classification. It was also observed that accuracy of co-registering temporal datasets has got a significant influence on the classification accuracy. However, the classification results because of mis-registration across data sets are quite complex to interpret. To address the cloud cover problem, additional training sites should be defined in cloudy regions keeping its temporal dynamics in view. Tthe overall classification accuracy in these cloud infested regions is visually estimated around 60-70%. Further, the accuracy a land cover class in cloud-infested regions depends on the extent of in the satellite image that provides major contribution to the signature of that class.

REFERENCES

Carpenter, G. A., Gopal, S., Macomber, S., Martens, S., Woodcock, C. E. and Franklin, J., 1999. A neural network method for efficient vegetation mapping. *Remote Sensing of Environment*, 70, pp. 326– 338.

Chavez, P.S. Jr., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment*, 24, pp. 459-479.

Friedl, M. A. and Brodley, C. E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61, pp. 399– 409.

Friedl, M. A., Muchoney, D., McIver, D., Gao, F., Hodges, J. F. C. and Strahler, A. H., 2000. Characterization of North American land cover from NOAA-AVHRR data using the EOS MODIS land cover classification algorithm. *Geophysical Research Letters*, 27 (7), pp. 977– 980.

Foody, G. M., 1997. Fully fuzzy supervised classification of land coverfrom remotely sensed imagery with an artificial neural network. *NeuralComputing and Applications*, 5 (4), pp. 238– 247.

Gopal, S., Woodcock, C. E. and Strahler, A. H., 1999. Fuzzy neural network classification of global land cover from a 1 degree AVHRR data set. *Remote Sensing of Environment*, 67, pp. 230–243.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G. and Sohlberg, R., 2000.Global land cover classification at the 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21 (6), pp. 1331–1364.

Hansen, M., Dubayah, R., & DeFries, R. (1996). Classification trees: An alternative to traditional land cover classifiers. International Journal of Remote Sensing, 17, 1075–1081.

Liang, S., Fang, H. and Chen, M., 2001, Atmospheric Correction of Landsat ETM+ Land Surface Imagery-Part I: Methods. *IEEE Transactions on GeoSciences and Remote Sensing*, 39 (11), pp. 2490-2498.

Quinlan, J. R., 1993. *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann.

Richards, J. A., 1993. *Remote sensing digital image analysis: an introduction*. New York: Springer-Verlag.

Strahler, A. H., 1980. The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of Environment*, 10, pp.135– 163.

Swain, P. H., and Hauska, H., 1977. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronic*s, 3, pp. 142–147.

Xu, M., Pakorn, W., Pramod, K. V. and Arora M.K., 2005. Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97, pp. 322 – 336.

ACKNOWLEDGEMENTS

The authors wish to place on record their sincere thanks to Dr. K. Radhakrishnan, Director, NRSA for his encouragement and extending necessary facilities required for study.