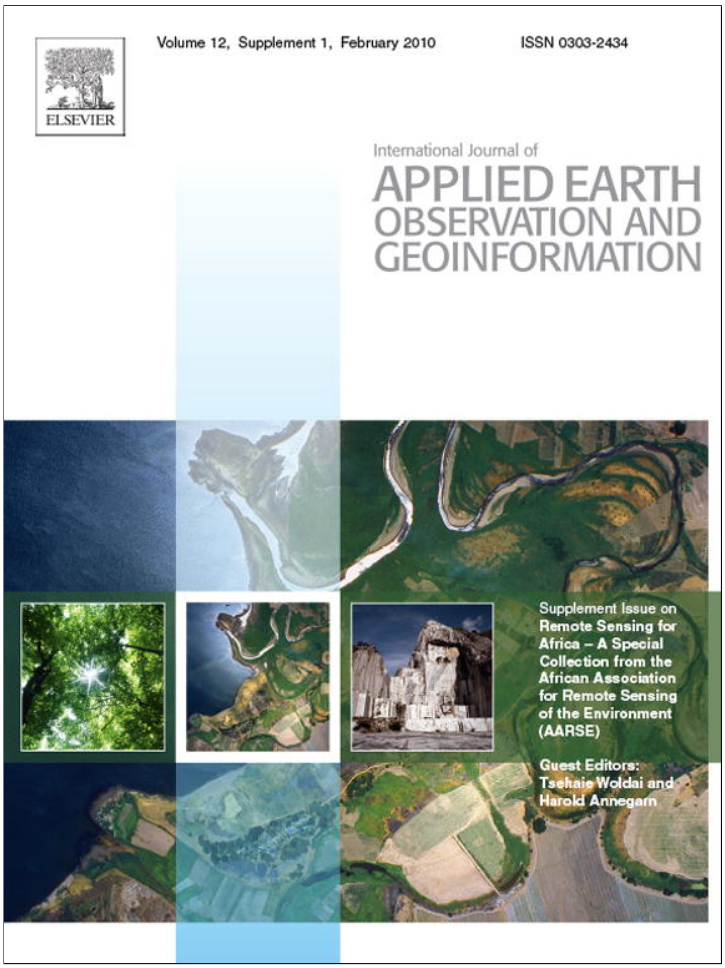

International Journal of Applied Earth Observation and Geoinformation 12S (2010) S27–S31

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms

J.R. Otukei [a,b,*], T. Blaschke [b]

[a] Department of Surveying, Makerere University, P.O. Box 7062, Makerere Hill Road, Kampala, Uganda
[b] Z_GIS Centre for GeoInformatics, University of Salzburg, Hellbrunner Str. 34, 5020 Salzburg, Austria

ARTICLE INFO

Article history:
Received 17 November 2008
Accepted 9 November 2009

Keywords:
Decision trees
Support vector machines
Maximum likelihood classifier
Land cover change

ABSTRACT

Land cover change assessment is one of the main applications of remote sensed data. A number of pixel based classification algorithms have been developed over the past years for the analysis of remotely sensed data. The most notable include the maximum likelihood classifier (MLC), support vector machines (SVMs) and the decision trees (DTs). The DTs in particular offer advantages not provided by other approaches. They are computationally fast and make no statistical assumptions regarding the distribution of data. The challenge to using DTs lies in the determination of the "best" tree structure and the decision boundaries. Recent developments in the field of data mining have however, provided an alternative for overcoming the above shortcomings. In this study, we analysed the potential of DTs as one technique for data mining for the analysis of the 1986 and 2001 Landsat TM and ETM+ datasets, respectively. The results were compared with those obtained using SVMs, and MLC. Overall, acceptable accuracies of over 85% were obtained in all the cases. In general, the DTs performed better than both MLC and SVMs.

1. Introduction

Land cover mapping and assessment is one of the core areas of remote sensing data application (King, 2002; Foody, 2002). Land cover is a fundamental variable that impacts on and links with many parts of the human and physical environment (Foody, 2002). The change in land cover is regarded as a single most important variable of global change affecting ecological systems (Vitousek, 1994) with an impact on the environment that is at least associated with climatic change (Skole, 1994). Despite the significant role that land cover information plays in environmental monitoring and understanding, our knowledge of land cover and its dynamics especially in the rural parts of Africa is still lacking.

The lack of knowledge relating to land cover and its dynamics especially in developing countries can be attributed to: (1) weak government support for mapping agencies and research institutions, (2) expensive software and hardware, (3) insufficient budget allocations for data purchases and (4) resistance to changes especially by the traditionalist in the field of mapping. However, the increasing availability of inexpensive or free data such as that provided by the global land cover facility (GLCF), the constant drop in the prices of hardware and software as well as improved awareness about the potential applications of remote sensing technology provides the needed momentum for land cover change assessment in the developing world. The combined use of remote sensing and geographical information systems (GIS) will render the essential tools for land cover mapping, storage, analysis and modelling of future scenarios (Geneletti and Gorte, 2003).

To effectively derive reliable information from satellite data, appropriate classification techniques are essential. A number of classification approaches have been developed over the past decades and a review of these algorithms can be found in Lu and Weng (2007). The classifiers can be categorised as either common or advanced. Some of the common classification algorithms include the K-Means, ISODATA, MLC and minimum distance to means (Erdas, 1999; Mather, 2004; Lillesand and Kiefer, 1999; Sabins, 1997; Richards, 1993) while the advanced classification algorithms include the artificial neural networks (ANN), decision trees, support vector machines, and object based image analysis (Lawrence et al., 2004; Mahesh and Mather, 2003; Kim et al., 2003; Mitra et al., 2004; Verbeke et al., 2004; Foody, 1986; Lucieer, 2008; Hay et al., 2003; Blaschke and Lang, 2006).

In this study, we explore the use of the DTs, SVMs and MLC approaches for land cover mapping as well as assessing the land cover changes in the rural areas of Pallisa District, Eastern Uganda. To this end, we pursue three main objectives: (1) explore the potential of data mining approaches for identification of suitable bands for classification as well as determining the decision thresholds, (2) compare the performance of the DTs, SVM and MLC and (3) assess the land cover changes within the study area over the given period.

* Corresponding author.
E-mail address: jrotukei@yahoo.com (J.R. Otukei).

0303-2434/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jag.2009.11.002

It is well established that the selection of a suitable classifier as well as appropriate bands (original or derived) is essential for improved classification accuracies (Lu and Weng, 2007). Consequently, the decision thresholds used for classification have an effect on the final outcome of the classification. Traditionally, the thresholds are obtained using the knowledge provided by experts who employ their expert knowledge to assess and create the decision boundaries. However, expert knowledge required to determine the decision boundaries is often lacking and this provides challenges for image classification. We argue that the use of a data mining approaches such as those implemented in WEKA software eliminates the burden of looking for expert knowledge for classification. Furthermore, since experts may disagree on the decision boundaries and it is difficult to know why they disagreed in the first place, we hypothesize that data mining approaches provide decision thresholds that are reliable, transferable and reproducible.

2. Study area

The study area is located in Kibale sub-county in Pallisa District, Eastern Uganda (Fig. 1). The geographical co-ordinates of the area of interest are: lat (1°11′–1°13′) N and long (33°44′–33°48′) E with maximum average elevation of approximately 1056 m a.s.l. The area is characterised by mostly savannah vegetation coupled with subsistence farmlands. The common food crops are millet, cassava, sorghum, potatoes as well as rice while cotton constitutes the main cash crop. Pallisa is one of the areas in Uganda known for having a large number of wetlands and as such it is an important area for conservation of the wetland habitat. However, over the last three decades, Pallisa has experienced land cover changes particularly as local people search for vacant land for cultivation. Most wetlands have been converted into rice gardens. Therefore assessing the land cover dynamics in this area is essential for understanding human interaction with their environment. This will form the basis for better planning and management of the existing resources.

3. Materials and methods

3.1. Data acquisition

Geo-referenced Landsat satellite images were accessed through the global land cover facility, courtesy of the NASA Landsat program. Only two datasets were available for the selected study area, i.e. the Landsat 5 (TM) for 1986 and Landsat 7 (ETM+) for 2001. These data sets are located on the satellite path 171 and row 59. The satellite images were acquired during late November and early January, respectively. It is generally a dry season in Pallisa district and therefore no significant spectral differences images are expected due to seasonal differences.

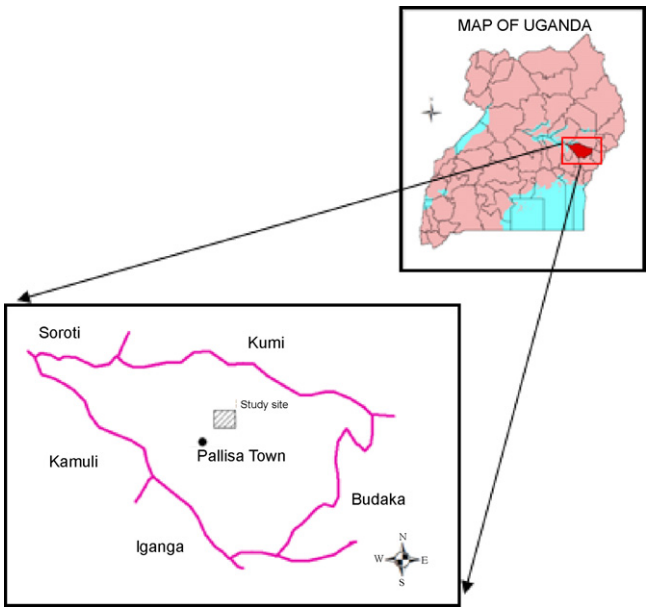

Fig. 1. Location of study area.

3.2. Data pre-processing

The data processing was carried out using ENVI 4.5 and ERDAS IMAGINE 9.1 prior to analysis. After the initial visual image analysis to confirm the agreement of the geo-referenced images, a subset of the image was extracted to include the area of interest and the surrounding areas. Apart from the original Landsat TM bands, a number of derivative bands were generated from the original data for analysis. These included the first three principle components (PCs), the first three tasselled cap (TC) transformed bands, normalised vegetation index (NDVI) and texture band based on spectral variance with a 3×3 moving window. The analysis of PCs showed that the first three PCs in both cases contained approximately 97% of the scene information and the remainder of the components with approximately 3% of the scene variance were not used for analysis. The three TC transformed bands were chosen since they represent the "*greenness*", "*brightness*" and "*wetness*" axes, respectively and therefore provide a measure of the presence or absence of vegetation as well as areas with high moisture content (Erdas, 1999). Similarly, the inclusion of the NDVI provides a measure of the absence and presence of vegetation. The NDVI has been used for vegetation studies especially assessing the health of vegetation (Morawitz et al., 2005), with higher NDVI values indicating good healthy vegetation while lower NDVI values show deprived vegetation. The original and derived bands were combined into a single band composite for land cover mapping.

3.3. Image classification

Image classification was performed using DTs, MLC and SVMs. In the following subsections a brief explanation of the three algorithms is provided.

3.3.1. Decision trees (DTs)

A decision tree classifier is a non-parametric classifier that does not require any *a priori* statistical assumptions to be made regarding the distribution of data. The process of building the decision tree is presented in Quinlan (1993). The basic structure of the decision tree however, consists of one root node, a number of internal nodes and finally a set of terminal nodes. The data is recursively divided down the decision tree according to the defined classification framework. At each node, a decision rule is required and this can be implemented using a splitting test often of the form

$$\sum_{i}^{n} a_i x_i \leq c \quad \text{for multivariate decision trees or simply} \quad x_i > c$$

for univariate decision trees.

where x_i represents the measurement vectors on the n selected features and a is a vector of linear discriminate coefficients while c is the decision threshold (Brodley and Utgoff, 1992). The DTs are known to produce results of higher accuracies in comparison to traditional approaches such as the "box" and "minimum distance to means" classifiers but the performance of DTs can be affected by a number of factors including: pruning and boosting methods used and decision thresholds (Mahesh and Mather, 2003). Our study addressed the challenges of determining the decision thresholds using data mining approaches.

3.3.2. Maximum likelihood classification

A maximum likelihood classification algorithm is one of the well known parametric classifies used for supervised classification. According to Erdas (1999) the algorithm for computing the weighted distance or likelihood D of unknown measurement vector X belong to one of the known classes M_c is based on the Bayesian equation.

$$D = \ln(a_c) - [0.5\ln(|cov_c|)] - [0.5(X - M_c)T(cov_c - 1)(X - M_c)]$$

The unknown measurement vector is assigned to the class in which it has the highest probability of belonging. The advantage of the MLC as a parametric classifier is that it takes into account the variance–covariance within the class distributions and for normally distributed data, the MLC performs better than the other known parametric classifies (Erdas, 1999). However, for data with a non-normal distribution, the results may be unsatisfactory.

3.3.3. Support vector machines

The support vector machines (SVMs) are a set of related learning algorithms used for classification and regression. Like the DTs classifiers, the SVM are also non-parametric classifiers. The theory of the SVM was originally proposed by Vapnik and Chervonenkis (1971) and later discussed in detail by Vapnik (1999). The success of the SVM depends on how well the process is trained. The easiest way to train the SVM is by using linearly separable classes. According to Osuna et al. (1997) if the training data with k number of samples is represented as $\{X_i, y_i\}$, $i = 1, \ldots, k$ where $X \in R^N$ is an N-dimensional space and $y \in \{-1, +1\}$ is a class label then these classes are considered linearly separable if there exists a vector W perpendicular to the linear hyper-plane (which determines the direction of the discriminating plane) and a scalar b showing the offset of the discriminating hyper-plane from the origin. For the two classes, i.e. class 1 represented as -1 and class 2 represented as +1, two hyper-planes can be used to discriminate the data points in the respective classes. These are expressed as

$$WX_i + b \geq +1 \text{ for all } y = +1, \text{ i.e. a member of class 1}$$
$$WX_i + b \leq -1 \text{ for all } y = -1, \text{ i.e. a member of class 2}$$

The two hyper-planes are selected so as not only to maximise the distance between the two given classes but also not to include any points between them. The overall goal is to find out in which class the new data points fall. Overall, the SVMs are reported to produce results of higher accuracies compared with the traditional approaches but the outcome depends on: the kernel used, choice of parameters for the chosen kernel and the method used to generated SVM (Huang et al., 2002).

3.3.4. Classification scheme

The first step in the classification process was the development of the classification scheme. The land cover classification scheme consisting of eight main land cover classes (mixed forest, degraded forest, herbaceous wetlands, shrub wetlands, grassland, grassland (open), mixed farmland and water (open)) were developed based on the Afri-cover land cover classification system (FAO, 2005). Also prior to digital image classification, appropriate bands for classification were determined using the C 4.5 data mining algorithm (Quinlan, 1993). Additionally, the decision rules for implementation using DT were generated using the same algorithm. After the determination of the appropriate bands, a classification was performed using MLC and SVM. In order to minimise the biasness caused by using different band combinations, the same number of bands were used for classification in both cases.

3.3.5. Pre-processing of training and test data

Training and test data for the associated classes were delineated based on analyst's prior knowledge of the study area. Further pre-processing of the training data was performed prior to analysis using data miner. Overall, 27 instances/columns were available for analysis, i.e. the 19 original and derived bands, land cover/land use class definition, 6 location variables for each pixel{pixel (x, y), map (lat, long), and map (X, Y)} and finally the identifier (ID) for each training data point. For this study, all the location variables as well as ID were not used for classification. As a result, the 19 original and derived bands were used as the dependent variables whereas the land cover classes were used as the independent variables.

The DT classifiers, i.e. number of bands and decision threshold for each image dataset were developed using the training dataset, by implementing the C4.5 data mining algorithm developed by Quinlan (1993). For each land cover class, at least 50 instances/pixels were available for analysis. A total of 768 and 643 instances were used for the Landsat TM and ETM+ data, respectively. In order to reduce the complexity of the tree classifies, pruning was enforced using a confidence factor setting of 0.25. Each of the resultant DTs were tested using the 10-fold classification approach. Accuracies of 96.6% and 95.8% were obtained for the TM and ETM+ datasets, respectively. The data mining approach provided two significant results, i.e. the rules as well as the appropriate bands for classification. Altogether, 11 and 10 bands (original and derived) were found appropriate for the classification of the 2001 and 1986 Landsat data, respectively. Whereas these bands where appropriate, only the Landsat TM bands 3 and 4 where appropriate for discriminating all the classes of interest.

3.3.6. Classification

The resultant tree classifiers were used for image classification using ENVI 4.4 software. Furthermore, the appropriate bands derived using C4.5 data mining algorithm were used for classification using the MLC and SVMs. The SVM classification was performed using a well known radial basis function kernel. For the MLC and SVM approach, the classification was performed in two stages. In the first stage, the same numbers of bands used for DT classification were also used for MLC and SVM classification. However, further analysis showed that some bands were not very significant in the DT classification. These bands were excluded in the second classification stage. The motive was to further evaluate whether these excluded bands had any significant impact on the accuracy of the classification. The second classification approach was not possible for the case of DTs classification. The results of the classification were smoothed with a 3 by 3 majority filter to minimise the salt and pepper appearance. The final results were used for accuracy assessment based on the confusion matrix. A separate data set not used for training was used for accuracy assessment.

4. Results

4.1. Classification results

Ten land cover classification results were obtained. The six results shown in Fig. 2 correspond to the three classifications of the 1986 and 2001 Landsat data based on the three classification algorithms. The additional four classifications refer to the classifications using SVM and MLC for both years based on the modified number of bands discussed in Section 3.3.6. In the 1986 satellite image, seven classes were successfully delineated (i.e. *forest*, *herbaceous wetlands*, *shrub wetlands*, *grassland*, *grassland* (*open*), *mixed farmland* and *water* (*open*)). However, it was not possible to identify the class '*Water (open)*' in the 2001 satellite image. Also the original class '*Forest*' was replaced by a new class

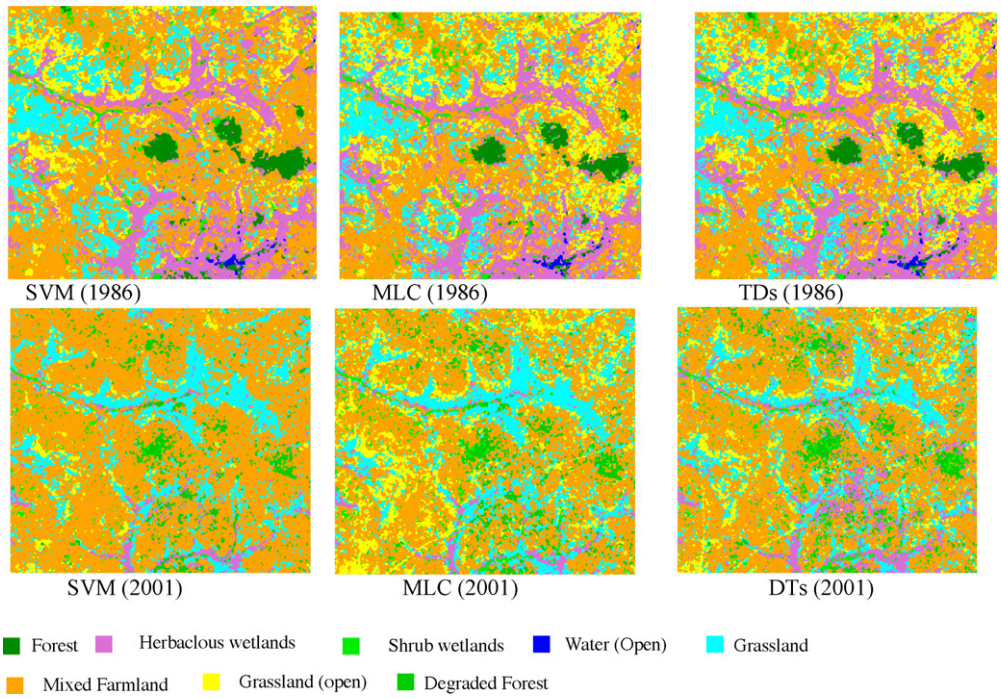

Fig. 2. Land classification with SVMs, MLC and DTs.

'*Degraded forest*'. Results show that the study area is dominated by subsistence "*Mixed Farmlands*".

4.2. Accuracy assessment

The classification accuracy was evaluated using the confusion matrix. A separate but same data set was used for accuracy assessment in all cases. The subscripts 1 and 2 in Table 1 refer to SVM and MLC classifications based on the original appropriate bands from the data mining algorithms and the modified classification after further reduction of bands, respectively. The first five columns show the results of the accuracy assessment of the 1986 imagery whereas the last five columns are for the 2001 satellite imagery.

4.3. Land cover change assessment

Fig. 3 shows the land cover changes which are identified within the study area. We adopted a post-classification approach for land cover change assessment. It is however important to note that each of the three adopted image processing techniques gave a different area estimate. This is not surprising since the choice of the classification techniques has an impact on the area estimate. But since each of the approaches provided results of acceptable accuracies, we considered all the resultant area estimates for land cover change assessment. In our approach, we simply averaged the area estimates from each of the classification techniques for each year. The results were subtracted to provide the change statistic. The results show an increase of the subsistence *mixed farmlands*, *grassland* and *degraded forest* while a decrease in *grassland* (*open*) and *herbaceous wetland*. The class *water* (*open*) and *forest* do not appear in the 2001 satellite image.

5. Discussion

The study area considered here was one of those areas in Eastern Uganda that were affected by insurgency for the period 1986–1992. Prior to this period, there were three small forested areas within the sub-county head quarters of Kibale. These forests were still identifiable in the 1986 satellite image. These were the *Okeju*, *Otelepai* and *Kibale* forests. *Okeju* was particularly known as an area where people were often abducted and murdered. To date, these forests no more exist. The classification results in the 2001 satellite data confirmed this observation. The forests have been converted to subsistence *mixed farmlands* due to the growing population and need for arable land. However, during the fallow periods these former forested areas tend to regenerate with some characteristic features of the original forest (often tall grass and bushes), which may appear as forested areas in the satellite image.

It is evident that the open waters have disappeared and there is also a general decrease in the herbaceous wetlands. Most of these areas have been converted into rice fields. December–February is always the rice clearing and growing season in Pallisa. There is, therefore, a possibility of these fields appearing as open grassland (rice growing at an early stage dominated with bare ground) or farmlands and grassland (i.e. when the fields are fully covered with rice). The results of the accuracy assessment showed high confusion between grassland, grassland (open) and herbaceous wetlands.

Outside the wetland areas, most grassland areas have been converted into farmlands. Overall, the wetlands and forest are disappearing while the farmlands are increasing. There have been some unsuccessful attempts to gazette wetlands in this area and other parts of Pallisa. The wetlands provide potential for rice growing which in fact has become both food and cash crop.

Table 1
Classification accuracy.

Method	DTs	SVM_1	SVM_2	MLC_1	MLC_2	DTs	SVM_1	SVM_2	MLC_1	MLC_2
Overall accuracy	93.48	90.53	89.49	90.42	87.30	94.07	91.73	93.67	93.91	93.67
Kappa statistics	0.93	0.89	0.87	0.88	0.85	0.93	0.90	0.92	0.93	0.92

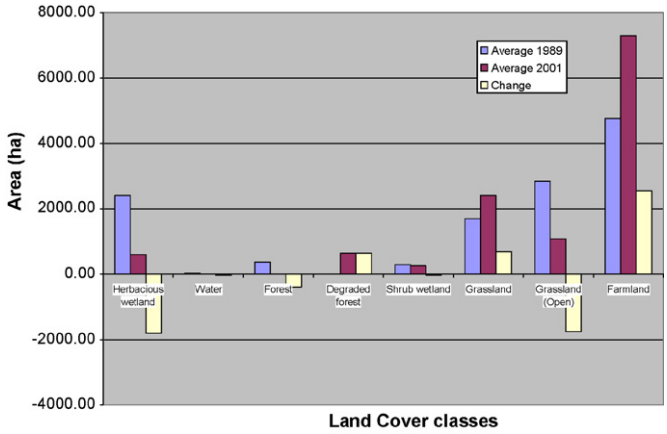

Fig. 3. Land cover changes.

Coupled with lack of land for most growing adults, it is envisaged that this trend will continue unless alternative sources of income are provided to the people.

Regarding the classification approach used, it is evident that data mining approaches when combined with traditional digital image classification provides potential for mapping and understanding environmental changes. The use of suitable data miners helps in choosing the appropriate threshold for classification as well as the bands for analysis. This eliminates the trial and error methods often used especially when classifying data of high dimensionality.

High overall accuracies were obtained for all the three techniques. However, the DTs performed better in both cases. The reduction of bands by eliminating the less appropriate bands does not significantly decrease the accuracy of the classification as can be observed in Table 1. Indeed for the case of SVM, there is an improvement of the classification accuracy. This is perhaps due to the simplification of the vector space needed for the development of hyper-planes.

6. Conclusions

The main aim of this study was to explore the data mining approaches for pixel based land cover classification as well as assessing the land cover changes that have occurred in a given area using the Landsat data of 1986 and 2001. The data mining approach provides a lot of potential for pixel based classification when used in conjunction with the traditional digital image classification processes. In particular, the ability to enable the identification of appropriate bands for classification and the determination of decision thresholds is plausible. This methodology provides results that are reliable, reproducible and transferable. The accuracies obtained are high and therefore considered acceptable. Also, the approach of determining land cover changes using results from different methods is commendable since it is not biased to a particular method. The study also concludes that land cover dynamics is occurring at an unprecedented rate.

Acknowledgements

The authors would like to thank the global land cover facility for the data that was accessed through the Internet free of charge. Furthermore, we would like to thank the University of Waikato for availing the open source data mining software WEKA. Finally, we would also like to express our gratitude to Peter Zeil, Centre for GeoInformatics, for his comments and suggestions during the preparation of this manuscript.

References

Blaschke, T., Lang, S., 2006. Object based analysis for automated information extraction-a synthesis. In: MAPPS/ASPRS Fall Conference, San Antonio, TX.
Brodley, C.E., Utgoff, P.E., 1992. Multivariate versus univariate decision trees. Technical Report 92-8. University of Massachusetts, Amherst, MA, USA.
Erdas Inc., 1999. Erdas Field Guide. Erdas Inc., Atlanta, Georgia.
FAO, 2005. Land Cover Classification System (LCCS), Classification Concepts and Users Manual. FAO, Rome, Italy.
Foody, G.M., 1986. Approaches for the production and evaluation of fuzzy land cover classification from remotely sensed data. International Journal of Remote Sensing 17, 1317–1340.
Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of the Environment 80, 185–201.
Geneletti, D., Gorte, B.G.H., 2003. A method for object oriented land cover classification combining landsat TM and aerial photographs. International Journal of Remote Sensing 24, 1273–1286.
Hay, G., Blaschke, T., et al., 2003. A comparison of three image-object methods for the multiscale analysis of the landscape structure. ISPRS Journal of Photogrammetry and Remote Sensing 57, 327–345.
Huang, C., Davis, L.S., Townshed, J.R.G., 2002. An assessment of support Vector Machines for Land cover classification. International Journal of Remote sensing 23, 725–749.
Kim, H., Pang, S., et al., 2003. Automatic land cover analysis for Tenerife by supervised classification using remote sensing data. Remote Sensing of the Environment 86, 530–541.
King, R.B., 2002. Land cover mapping principles: a return to interpretation fundamentals. International Journal of Remote Sensing 23, 3525–3546.
Lawrence, R., Bunn, A., et al., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sensing of the Environment 90, 331–336.
Lillesand, T.M., Kiefer, R.W., 1999. Remote Sensing and Image Interpretation. John Wiley and Sons Ltd., New York.
Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 26 (5), 823–870.
Lucieer, V., 2008. Object-oriented classification of sidescan sonar for mapping benthic marine. International Journal of Remote Sensing 29 (3), 905–921.
Mahesh, P., Mather, P.M., 2003. An assessment of the effectiveness of the decision tree method for land cover classification. Remote Sensing of the Environment 86, 554–565.
Mather, P.M., 2004. Computer Processing of Remotely Sensed Images: An Introduction. John Willey and Sons Ltd., West Sussex, England.
Mitra, P., Shankar, B.U., et al., 2004. Segmentation of multi-spectral remote sensing images using active support vector machines. Pattern Recognition Letters 25, 1067–1074.
Morawitz, D.F., Blewett, T.A., et al., 2005. Using NDVI to assess vegetative land cover in central Puget sound. Environmental Monitoring and Assessment 114, 85–106.
Osuna, E.E., Freud, R., et al., 1997. Support Vector Machines: Training and Applications, A.I. Memo No. 1602, C.B.C.L. Paper No. 144. Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.
Quinlan, R., 1993. Programs for Machine Learning. Morgan Kaufman, San Mateo.
Richards, J.A., 1993. Remote Sensing Digital Image Analysis. An Introduction. Springer-Verlag, Berlin.
Sabins, F.F., 1997. Remote Sensing: Principles and Interpretation. W.H. Freeman and Company, New York.
Skole, D.L., 1994. Data on global land cover change: acquisition assessment and analysis. In: Turner, II, W.B. (Ed.), Changes in Land Use and Land Cover: A Global Perspective. Cambridge University Press, Cambridge, pp. 437–471.
Vapnik, W.N., 1999. An overview of statistical learning theory. IEEE Transactions of Neural Networks 10, 988–999.
Vapnik, W.N., Chervonenkis, A.Y., 1971. On the uniform convergence of the relative frequencies of events to their probabilities. Theory of Probability and its Applications 17, 264–280.
Verbeke, L.P.C., Vabcoillie, F.M.B., et al., 2004. Re-using back propagating artificial neural network for land cover classification in tropical savannahs. International Journal of Remote Sensing 35, 2747–2771.
Vitousek, P.M., 1994. Beyond global warming: ecology and global change. Ecology 75, 1861–1876.