Research Journal of Applied Sciences, Engineering and Technology 6(15): 2881-2885, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: April 06, 2013 Accepted: April 22, 2013 Published: August 20, 2013

K-Nearest Neighbor Method for Classification of Forest Encroachment by Using Reflectance Processing of Remote Sensing Spectroradiometer Data

[1]Ahmed A. Mehdawi and [2]Bahrin Bin Ahmad
[1]UTM, Skudi, Johor, Malaysia
[2]Faculty of Geoinformation and Real Estate, Institute of Geospatial Science and Technology (INSTeG), Universiti Teknologi, Malaysia

Abstract: This study gives sophisticated result in the use of K-Nearest Neighbor Method classification of forest. The major focus is on the data and technique that can be used to identify the changes in forest features. This study will concentrate on identifying forest encroachment in tropical forests such as the forests of Malaysia. This technique study will establish a strong mechanism that can be used by different sectors such as forestry, local administration, surveying and agriculture. The main contribution of this study is that it utilizes of K-Nearest Neighbor Method with remote sensing data to detect forest encroachment. Hopefully, this study will serve as a reference for any future research on utilizes of K-Nearest classification as tools to identify of tropical forest encroachment.

Keywords: K-nearest neighbor method, malaysia and encroachment, remote sensing, tropical forests

INTRODUCTION

Preservation farm land utilizes arranging within humid tropical lowland forests and appropriate remote sensing approaches to differentiate amongst the many floristically different forest types. Satellite images and airborne photographs would definitely be a most important provider of greenery data in remote tropical zones in which various other data sources, for instance maps of vegetation, soil along with topography are often inaccessible (Ferrier, 2002). Floristically various tropical rain forest zones repeatedly lack wide-ranging field data as a result of logistical difficulties as well as challenges with outlining incredibly diverse or perhaps even negatively known flora (Ruokolainen *et al*., 1997). Whenever associated with species makeup or structural circumstances on the ground, satellite images produce a low-cost source most typically associated with data for the purpose of determining and mapping rain forest kinds but for some data only, a necessary with respect to preservation planning and maintainable forest management (Favrichon, 1998; Ferrier and Guisan, 2006; Margules and Pressey, 2000).

Tropical rain forest kinds are likely to be categorized through extended physiognomic and design characteristics, mainly because they normally seem structurally comparatively homogeneous within terrain dimensions. On top of that, the field details are in most cases as well coarse just for identifying floristic variance as a result of selecting mistake coming from variations which are usually as a consequence of certain geographical elements. Accordingly, just a couple of classes are actually employed to differentiate rain forest kinds making use of remote sensing methods (Achard *et al*., 2001; Kleinn *et al*., 2002; Stibig *et al*., 2003). Lately, numerous researches have shown the fact that rain forests have a relatively wide range of soil-related fine-grained floristic and architectural variation, that isn't viewable through the existing greenery maps.

The quality of natural environment classification accuracy and reliability originating from a satellite data is determined by the classification algorithm (Kleinn *et al*., 2002) as well as the image resolution (pixel window or segment size) utilized for the process. Additionally, distinctly identified forest classes are required to assess classifiers to get thematic accuracy. A commonly utilized supervised classification strategy is discriminant analysis (Thenkabail *et al*., 2004).

Discriminant analysis mission to find the linear mix off variables (e.g., spectral features) that most effective discriminates within classes. A non-parametric replacement for discriminant analysis is the K nearest neighbours (K-NN) classifier.

The K-NN technique has usually been employed to calculate forest inventory variables from satellite imagery, such as total volume and basal area for temperate and boreal zones (Gjertsen *et al*., 2000; McRoberts *et al*., 2007), on the other hand, it happens to be significantly less common with regard to image relying classification of forest kinds. Within the tropics it has been examined once, with promising outcomes, to get estimating frequent floristic dissimilarities. A particular advantage of the non-parametric K-NN classifier is that it will not make any distributional assumptions in regards to the variables used. In K-NN, the pixel whose class is unfamiliar is a member of a

Corresponding Author: Ahmed A. Mehdawi, UTM, Skudi, Johor, Malaysia

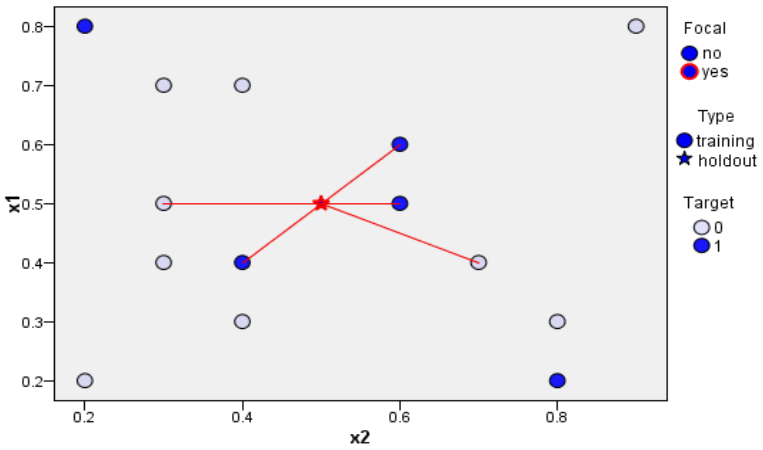

Fig. 1: Shown distance between the holdout point (star) and the training point (blue/white)

Fig. 2: Showed QuickBird Image for study area in 2007

class as outlined by its spectrally nearest neighbours (e.g., spectrally most equivalent pixels) whose class identities are recognized.

The K-Nearest Neighbor (KNN) multisource inventory has turned out to be timely, cost-efficient, as well as precise while in the tropical forest and Malaysian trials. This strategy designed for improving field point inventories will be perfectly appropriate for the evaluation and observation requires of Authorities agencies, such as the Forest Service, that carry out natural and additionally agricultural resource inventories. It includes wall-towall maps of forest features, continues the natural data variety perfectly found on the field inventory and presents accurate and localized estimates in accordance metrics throughout significant areas as well as ownerships.

Within a pixel-level classification, the KNN algorithm assigns every unidentified (target) pixel the field features of essentially the most equivalent reference pixels for which field data occurs (Franco-Lopez *et al*., 2001). Likeness is defined in conditions with the feature space, traditionally measured as Euclidean or Mahalanobis as the equation 1 distance between spectral featuresin addition Euclidean distance function is used for measuring the distance between the holdout point (star) and the training point (blue/white) as shown in Fig. 1 .Associated with principle purpose should be to explore whether discriminant analysis and k-NN classifiers will be able to efficiently classify study area (Cameron Highland as rain forest types to identify the most suitable classification approach via accuracy to detect forest change Cases that are near each other are seemed to be "neighbours." When a new case (holdout) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases – the nearest neighbours – are tallied and the new case is placed into the category that contains the greatest number of nearest neighbors.

$$d_E(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (1)$$

In kNN classification of Forest Inventory and Analysis FIA-defined forest kinds. Particular importance is determined relating to increasing mapping productivity by lessening classification feature space, reducing the number of distance estimations in the most adjacent neighbor search, along with eliminating redundancy in redundant nearest neighbor searches by constructing a database of feature patterns connected with different forest kind classes.

Methods and dataset:

General information forthe study area: The district of Cameron Highlands (4°28' N) (101°23' E), Pahang, Malaysia, is located on the main range of Peninsular Malaysia as shown in Fig. 2 and 3. It covers a total area of 71 000 715 km² (Fortuin, 2006). Generally, the terrain is mountainous and strongly dissected with 10–35° slopes. More than 66 per cent of the land has a gradient of more than 20°. The Cameron Highlands are about 715 km² in area settled between roughly 900 and 1800 m and surrounded by forested peaks rising to 2032 m. Malaysian lowlands are heavily disturbed, so upland forests like those of the Cameron Highlands are an important refuge for biodiversityas shown in Fig. 2 by QuickBird satellite . The Cameron Highlands are significantly cooler than Malaysia's lowlands, with a mean daily minimum of 14.8ºC, a mean daily maximum of 21.1ºC, which suits temperate crops. The rainfall averages 2660 mm yr-1, humidity is high and there is no marked dry season.

Fig. 3: The location of Cameron Highland located between Perak and Pahang (Malaysian)

Table 1: Summary of the dataset

Class	Representation of class	Feature
Bush	1	Reflectance
Cover	2	Reflectance
Forest	3	Reflectance
Tea	4	Reflectance
Vege	5	Reflectance
Zink	6	Reflectance

Dataset: This dataset is categorized as multiclass dataset. The dataset consists of one feature (average of reflectance) and 6 classes. The classes are named as follows:

- Bush
- Cover
- Forest
- Tea
- Vege
- Zink

Each class consists of 2151 samples of total 12906 samples processed on Mata lab software. The summary of the dataset is tabulated in Table 1.

Methods: Raining and testing data is divided using the method of 3-fold cross validation. For experimenting using MATLAB, user requires to firstly choose the directory of m file as shown below in Fig. 4, beside that the flow chat of the method summarized the study steps as showed in Fig. 5.

With regard to calculating by using Euclidean distances, take into account the spectral distance $d_{pi,p}$,that is definitely calculated in the feature space through the focus on pixel p to every single reference pixel $p_i$by which the forest kind class is recognized. For every single pixel p, sort the k-nearest field plot pixels (from the feature space) just by $d_{pi,p} \leq \ldots \leq d_{pk,p}$. The imputed value with the pixel p will likely be expressed as an objective of the nearest units, every single such

```
load('/Users/postgrade/Downloads/knn/data.mat');
class(1,:) = find(data(:,3) == 1);
class(2,:) = find(data(:,3) == 2);
class(3,:) = find(data(:,3) == 3);
class(4,:) = find(data(:,3) == 4);
class(5,:) = find(data(:,3) == 5);
class(6,:) = find(data(:,3) == 6);
```

Fig. 4: Shows the bath at MATLAB software to the user to manage the large number of data

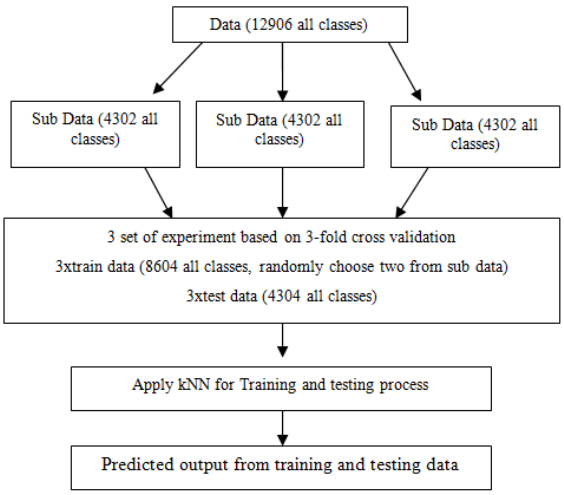

Fig. 5: Explain flowchart of the method s

Table 2: Description of MATLAB's Variables

No	MATLAB variable	Descriptions
1	Training data	Training data
2	Training class	Class of training data
3	Sample data	Testing data
4	Sample class	Class of testing data
5	Predicted class	Predicted class of testing data
6	Accuracy percentage	Accuracy from testing data
7	Test1-6	Testing data with their original output class
8	Testdata1-6	Testing data with their predicted output class

unit calculated as outlined by this distance breaking down function:

$$w_{pi,p} = \frac{1}{d_{pi,p}^{t}} / \sum_{i=1}^{k} \frac{1}{d_{pj,p}^{t}} \quad (2)$$

In which t can be described as distance decomposition factor set equivalent to 1 for all trials. To help impute class variables which include forest type and MATLAB variables as in Table 2, the distance decomposition function calculates a weighted mode value.

To get a class variable, the error rate (*Err*) suggests the difference of opinion between an estimated value $\hat{y}$and the real response$y$ in a dichotomous circumstance these types of which y does indeed or doesn't necessarily participate in class i (Efron and Tibshirani 1993). Therefore, by implementing the overall accuracy (OA) (Stehman 1997,) described as follows:

$$OA = 1\text{-}Err, \quad (3)$$

$$Err = \sum_{i=1}^{n} (y_i - \hat{y}) / n \quad (4)$$

Here is the exclusive circumstance with the mean square error for a great signal variable. A lot of these estimators have been recommended over the typical Kappa estimator regarding causes offered by Franco-Lopez *et al*. (2001).

Errors have been calculated by means of leave-one-out cross-validation. This method omits training sample models individually together with mimics the application of independent data .Per omission; we applied the kNN prediction principle towards the outstanding small sample. Eventually, the errors out there forecasts have been made clear. Overall, we employed the prediction rule n times and predicted the results to get n units.This kind of rates involving prediction error usually are practically unbiased (Efron and Tibshirani, 1993).

RESULT AND DISCUSSION

Simply because taken into account within the Methods portion, while k increases, a better probability is present which a reference observation is going to cross typically the inequality in addition to require the whole Euclidean distance contrast using the target. Our studies verified this unique relationship among increasing k as well as range of Euclidean distance measurements.

The majority of significantly, our studies show that using the *KNN* algorithm could substantially enhance mapping productivity by reducing the amount of measurements if necessary.

Typically the reduction within entire accuracy and reliability and the classification accuracy 99.9070 .Implementing our database-assisted mapping may possibly appreciate experience success whenever classes were much more spectrally distinct.

REFERENCES

Achard, F., H. Eva and P. Mayaux, 2001. Tropical forest mapping from coarse spatial resolution satellite data: Production and accuracy assessment issues. Int. J. Remote Sensing, 22: 2741-2762.

Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. Chapman and Hall, New York, pp: 436.

Favrichon, V., 1998. Modeling the dynamics and species composition of a tropical mixed-species uneven-aged natural forest: Effects of alternative cutting regimes. Forest Sci., 44: 113-124.

Ferrier, S., 2002. Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? Syst. Biol., 51: 331-363.

Ferrier, S. and A. Guisan, 2006. Spatial modelling of biodiversity at the community level. J. Appl. Ecol., 43: 393-404.

Fortuin, R., 2006. Soil erosion in cameron highlands, an erosion rate study of a highland area, Saxion University Deventer. Fragmentation and its impact on species diversity: An analysis using remote sensing and GIS. Biol. Conserv., 14: 1681-1698.

Franco-Lopez, H., A.R. Ek and M.E. Bauer, 2001. Estimation and mapping of forest stand density, volume and cover type using the k-nearest neighbors' method. Remote Sens. Environ., 77: 251-274.

Gjertsen, A.K., S. Tomter and E. Tomppo, 2000. Combined use of NFI sample plots Ad Landsat TM data to provide forest information on municipality level. Proceedings of Conference on Remote Sensing and Forest Monitoring, Rogow, Poland, pp: 167-174.

Kleinn, C., L. Corrales and D. Morales, 2002. Forest area in Costa Rica: A comparative study of tropical forest cover estimates over time. Environ. Monit. Assess., 73: 17-40.

Margules, C.R. and R.L. Pressey, 2000. Systematic conservation planning. Nature, 405: 243-253.

McRoberts, R.E., E.O. Tomppo, A.O. Finley and J. Heikkinen, 2007. Estimating areal means and variances of forest attributes using the k-nearest neighbors' technique and satellite imagery. Remote Sens. Environ., 111: 466-480.

Ruokolainen, K., A. Linna and H. Tuomisto, 1997. Use of melastomataceae and pteridophytes for revealing phytogeographical patterns in Amazonian rain forests. J. Trop. Ecol., 13: 243-256.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ., 62: 77-89.

Stibig, H.J., R. Beuchle and F. Achard, 2003. Mapping of the tropical forest cover of insular Southeast Asia from SPOT4-Vegetation images. Int. J. Remote Sens., 24: 3651-3662.

Thenkabail, P.S., E.A. Enclona, M.S. Ashton, C. Legg and M.J. De Dieu, 2004. Hyperion, IKONOS, ALI and ETM plus sensors in the study of African rainforests. Remote Sens. Environ., 90: 23-43.