Satellite Image Classification using Decision Tree, SVM and k-Nearest Neighbor

Article · July 2018

CITATIONS
0

READS
666

5 authors, including:

Iva Nurwauziyah
National Cheng Kung University
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Umroh Dian Sulistyah
National Cheng Kung University
4 PUBLICATIONS **1** CITATION
SEE PROFILE

I Gede Brawiswa Putra
National Cheng Kung University
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Muhammad Irsyadi Firdaus
National Cheng Kung University
12 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Satellite Image Classification using Decision Tree, SVM and k-Nearest Neighbor

Iva Nurwauziyah[1], Umroh Dian S.[2], I Gede Brawiswa Putra[3], Muhammad Irsyadi Firdaus[4]
[1,2,3,4] Department of Geomatics, National Cheng Kung University, Tainan, Taiwan
e-mail: ivanurwauziyah@gmail.com[1], umrohdians@gmail.com[2], brawiswa@gmail.com[3], irsyadifirdaus@gmail.com[4]

***Abstract* — There is undoubtedly a high demand for land use/cover maps for the monitoring and management of natural resources, development strategies, and global change studies. A variety of classification methods have been developed and tested for land use/cover to extract the knowledge. Knowledge discovery is also important because it gives a basic model for selection, preprocessing, transformation, data mining and interpretation of datasets. Here, the project are performed on satellite images with applying 3 different data mining classification method which is Decision Tree, SVM and k-NN methods to find the most efficient method. In geomatics, we used satellite image to measure the characteristics of an object or surface from a distance by interpretation of million number of pixel. It is now very complex and critical task for engineers and researchers to obtain the meaningful information from massive data sets. From those pixels image datasets, we can use data mining tools and efficient algorithms for getting meaningful information. In this project using Pleiades Satellite Images data in Taiwan. The result shows SVM method has best accuracy compared to the Decision Tree and k-Nearest Neighbor methods with the number of overall accuracy is 78.6% and 83.30% for high and low resolution satellite imagery respectively.**

I. INTRODUCTION

Image classification [7] is taken as growing field of both computer vision and data mining. We all know computer vision is the field of acquiring, processing, analyzing and understanding images which are later used for knowledge discovery from high-dimensional image data.

Remote sensing satellite images [3] are considered as one of the most important data sources for land use/cover mapping due to their extensive geographical coverage at an efficient cost while providing irreplaceable information on the earth's surface. Land use/cover maps are usually produced based on remote sensing image classification approaches. However, the accuracy and processing time of land use/cover maps using remote sensing images is still a challenge to the remote sensing community.

In order to understand a classification of satellite image, the first step is to recognize the objects and then recognize the category of the scene. In order to do this in computer vision, we use various classifiers that all have different characteristics and features.

Many classifiers have been developed by various researchers. These methods include naïve Bayes classifier, support vector machines, k-nearest neighbors, Gaussian mixture model, decision tree and radial basis function (RBF) classifiers.

In this paper, we will be comparing three different classification methods: Experimental evaluation is conducted on the high resolution satellite image and low resolution satellite image dataset to see the difference between three classification methods. Then, we will explain the three different classification performance of this three methods that we have used: DT, KNN and SVM.

II. RELATED WORK

According to Lu and Weng [5], it is not only the imagery appropriateness but also the right choice of classification method that affects the results of land use/cover mapping. In literature, a variety of classification methods have been developed and tested for land use/cover mapping using remote sensing data. These methods range from unsupervised algorithms (i.e., ISODATA or K-means) to parametric supervised algorithms (i.e., maximum likelihood) and machine learning algorithms such as artificial neural networks (ANN), k-Nearest Neighbors (kNN), decision trees (DT), support vector machines (SVM), and random forest (RF).

In [9], Landslide image data is taken for data mining purpose. Vegetation Index and the thresholds are of each attribute on target categories. A conventional approach, C4.5 Decision Tree Analysis, is used as a comparison. And it helps to analyze the landslide problems and thus facilitates the informed decision-making process.

In [4], comparison is based on traditional Classification tree results to stochastic gradient boosting for three remote sensing based data sets, an IKONOS image from the Sierra Nevada Mountains of California, a Probe-1 hyperspectral image from the Virginia City mining district of Montana, and a series of Landsat images from the Greater Yellowstone Ecosystem.

III. METHODOLOGY

A. Study Area

In this study, in order to compare the performance of different classification algorithms on different data training sample strategies, an area of 4300 x 4300 pixels of a peri-urban and rural with heterogeneous land cover area in the north of Taiwan was chosen (Figure 2). Second training data is High Resolution Satellite Image using Pleiades image with 1300 x 1300 pixels image size in colorado, USA was chosen (Figure 1). The study area mainly four typical classes: resident (fragment and distribution over the study

area), impervious surface (including factory, block building and transportation, roads), and park.

B. Image Dataset Used

Figure 1. *High Resolution Satellite Image*

Figure 2. *Low Resolution Satellite Image*

C. Flowchart

A support vector machine (SVM) [2], a promising method for classification of both linear and nonlinear data. SVM classification uses different planes in space to divide data points. It gains flexibility in the choice of threshold and handles more input data very efficiently. Its performance and accuracy depend upon the selection of hyper plane and kernel parameter. The goal of SVM Classification is to produce a model, based on the training data, which will be able to predict class labels of the test data accurately.

K-nearest neighbor (KNN) algorithm [1] is a method for classifying objects based on closest training examples in the feature space. *K*-nearest neighbor algorithm is among the simplest of all machine learning algorithms. Training process for this algorithm only consists of storing feature vectors and labels of the training images. In the classification process, the unlabeled query point is simply assigned to the label of its *k*-nearest neighbors.

The Decision Tree (DT) classifier [8] is one of the inductive learning algorithms that generates classification tree using the training data/sample. It is based on the "divide and conquer" strategy. It consists of mainly three parts: Partitioning the nodes, find the terminal nodes and allocate class label to terminal nodes. It is based on hierarchical rule. It handles high dimensional data and representation of knowledge in tree form which is easy to humans for understanding purpose. When decision tree built, many of branches reflects noise in the training pattern so, tree pruning attempts to identify and remove such branches and improve the accuracy of classification.

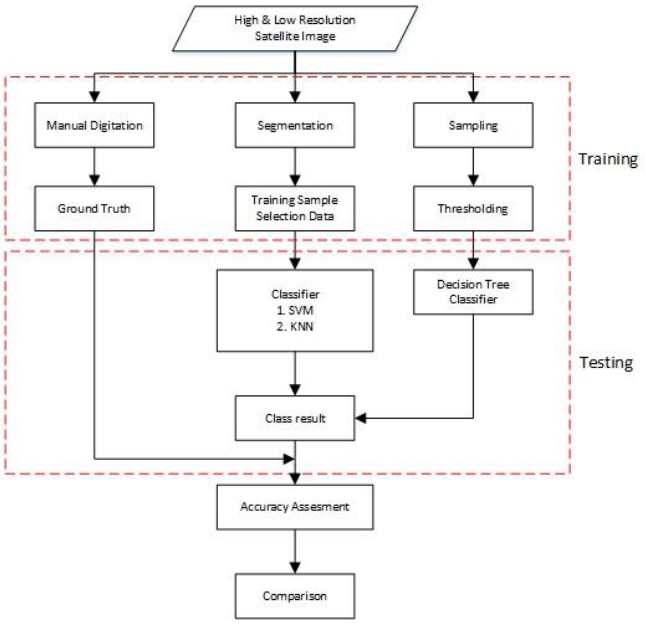

Figure 3. General Flowchart of SVM and K-NN classifier

The process is illustrated by the figure 3. Start from choosing high and low resolution that used for classification process. These images will be segmented based on the correlation between neighborhood pixels. Next step is selecting training sample data wherein this step needs to select 70% of the total pixel for generating a proper classification result. There are 2 methods that used in this project which are SVM and k-NN where in this stage, it need to select several parameters to run the algorithm. Lastly, the classification result is compared to the ground-truth data to get the accuracy assessment.

The DT classification algorithm is shown in figure 4 and 5 for high and low resolution satellite imagery, respectively. In the high resolution satellite imagery, B4 (NIR) is the band used for initial binary splitting of the

image, which is subsequently supported by B2 (green). While, in the low resolution satellite imagery, the only parameter that used to classified the images is NDVI.

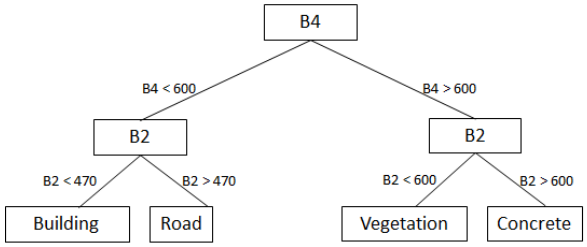

Figure 4. Thresholding of DT classifier for high resolution satellite image

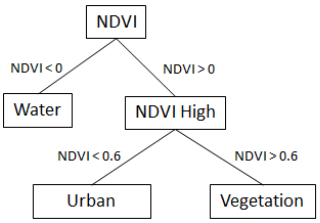

Figure 5. Thresholding of DT classifier for low resolution satellite image

Next, the classification result will compare with ground truth image which generate from manual interpretation and manual digitation to get the overall accuracy. Lastly, each accuracy will compare to judge the best classification method for high and low resolution images

D. Segmentation

Segmentation is the process of dividing an image into segments that have similar spectral, spatial, and/or texture characteristics. The segments in the image ideally correspond to real-world features. Effective segmentation ensures that your classification results are more accurate.

There two parameters that have to consider for segmentation which is scale level and merge level. Adjust the Scale Level slider as needed to effectively delineate the boundaries of features as much as possible without over-segmenting the features. While merging is used primarily to combine segments with similar spectral information

For high resolution satellite image, the scale level is set into 40 and merge level is 80. Similar to the low resolution, the scale and merge level are set into 40 and 90 respectively. If the Scale Value is higher, some rooftop segments would be combined with segments representing adjacent backyards or trees because they have a similar intensity.

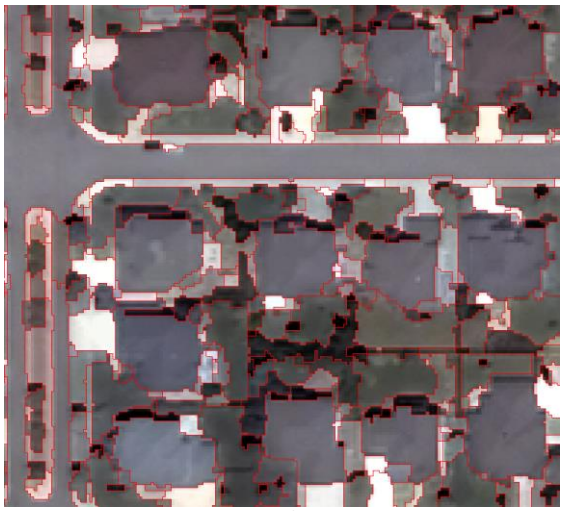

Figure 6. *Segmentation Result of High Resolution Satellite Image*

In the Pleiades image (figure 6), the rooftops appear much darker, while the concrete area is lighter. The road already segmented very well and the vegetation area has many tiny segment area because it has many different color in example tree and grass has different color which automatically will segmented.

Figure 7. *Segmentation Result of Low Resolution Satellite Image*

Segmentation on Landsat image (figure 7), show the urban area have smaller segment size than vegetation area, because urban area has various pixel value. However, the water region has least segment area, because the pixel value is very similar.

E. Training and Testing Sample Datasets

The training data (training and testing samples) was collected based on the manual interpretation of the original Pleiades data and low-resolution imagery available from Landsat 8. To collect training sample data, the create training sample in the ENVI 5.3

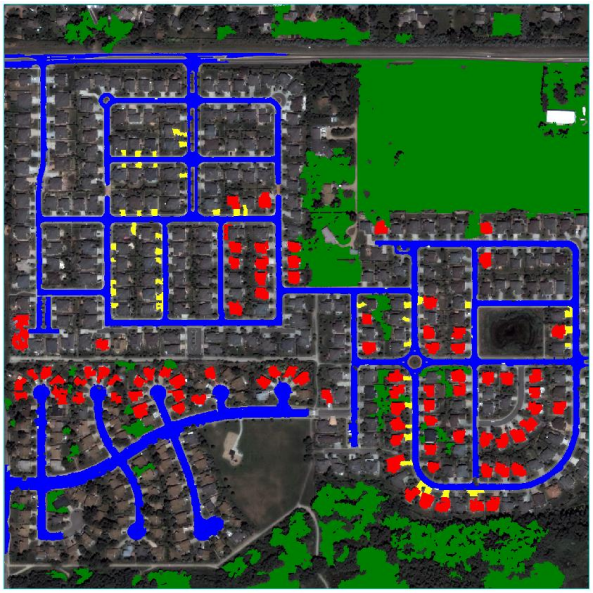

Figure 8. Training data for high resolution satellite image

The training data are obtained as shown in Fig. 8, and Fig. 9. Training data for high resolution satellite image (Fig. 8), the red color shows building region, the green color shows vegetation region, blue color shows road region and yellow color shows concrete region. While, training data for low resolution satellite image (Fig. 9), the red color shows urban region, the green color shows vegetation region and blue color shows water region.

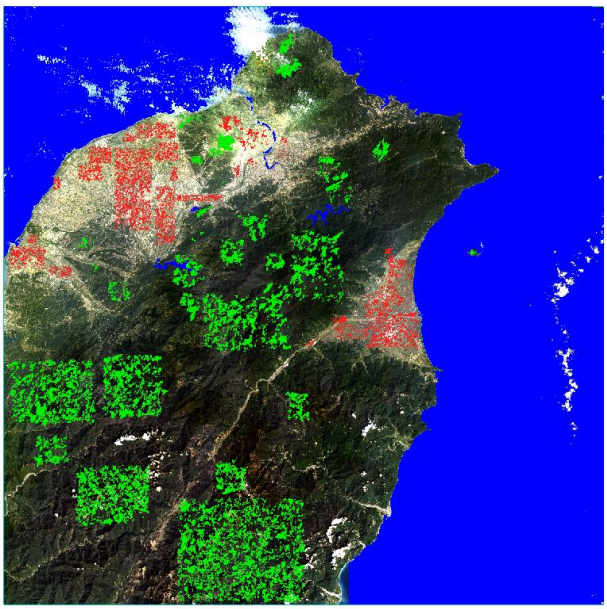

Figure 9. Training data for low resolution satellite image

Table 1. Training sample sizes of high resolution satellite image used in this study

Land Cover	Training Area (Segment)
Vegetation	1726
Building	486
Road	72
Concrete	420

Table 2. Training sample sizes of low resolution satellite image used in this study

Land Cover	Training Area (Segment)
Vegetation	11275
Water	98
Urban	10964

IV. RESULT AND DISCUSSION

All the experiments are carried out in Intel(R) Core(TM) i7 4.20 GHz processor with 1 TB HDD, 32 GB RAM, Windows 10 Operating system. Here, image analysis tool is used for our experiment with Data mining application in ENVI version 5.3.

A. Low Resolution Satellite Image

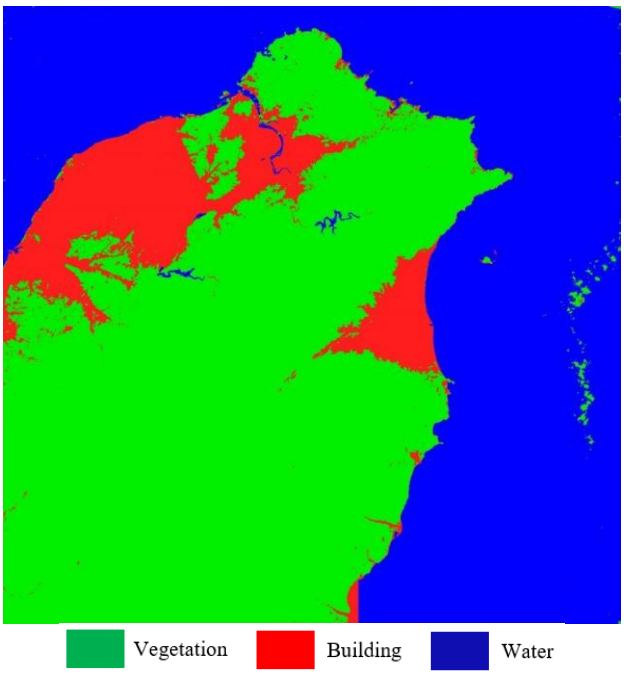

Figure 10. SVM Classification

Table 3 shows the confusion matrix performance of SVM classifier for each individual class with the actual class represented on the above and the predicted class represented on the left. the diagonally shaded boxes show

the percent of accurately classified images for each class while the other boxes show the percent of erroneously classified images.

Table 3. The Confusion Matrix Performance of SVM classifier

Class	Ground Truth (Percent)		
	Vegetation	Water	Urban
Vegetation	**97.22**	0.09	50.42
Water	0.19	**99.56**	10.14
Urban	2.59	0.35	**39.43**

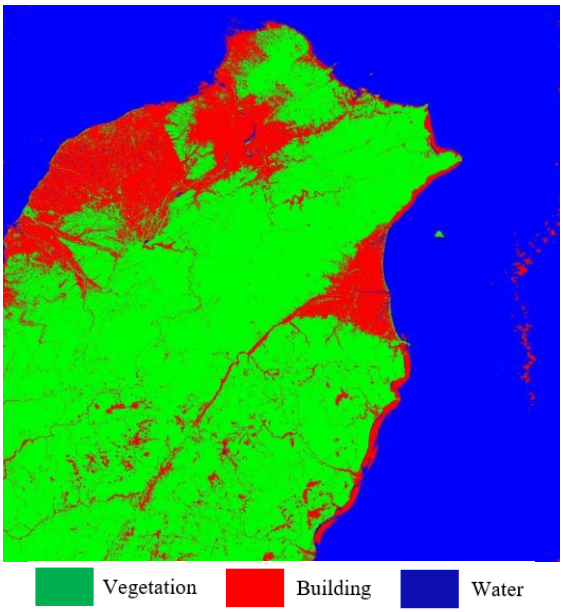

Figure 11. Decision Tree Classification

Vegetation as actual class and vegetation as predicted class has value confusion matrix is high when compared with other classes. It is indication

Table 4. The confusion matrix Performance of DT classifier

Class	Ground Truth (Percent)		
	Water	Urban	Vegetation
Water	**49.48**	8.36	0.07
Urban	1.02	**37.1**	9.25
Vegetation	0.02	46.18	**90.6**

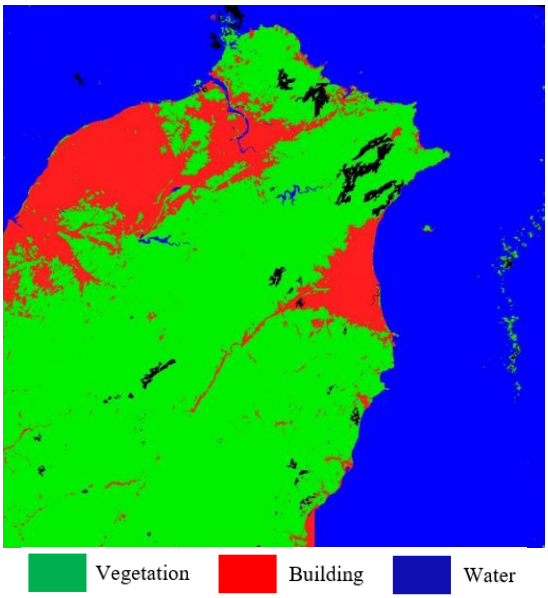

Figure 12. k-Nearest Neighbor

Table 5. The Confusion Matrix Performance of KNN classifier

Class	Ground Truth (Percent)		
	Vegetation	Water	Urban
Vegetation	**92.87**	0.07	46.53
Water	0.22	**99.54**	10.18
Urban	4.79	0.34	**42.53**

B. High Resolution Satellite Image

Figure 13. SVM Classification

Table 6. The Confusion Matrix Performance of SVM classifier

	Ground Truth (Percent)			
Class	Building	Road	Vegetation	Concrete
Building	**87.4**	1.36	16.82	0.46
Road	0.07	**91.47**	7.55	0.78
Vegetation	11.74	5.04	**74.59**	2.41
Concrete	0.79	2.14	1.04	**96.36**

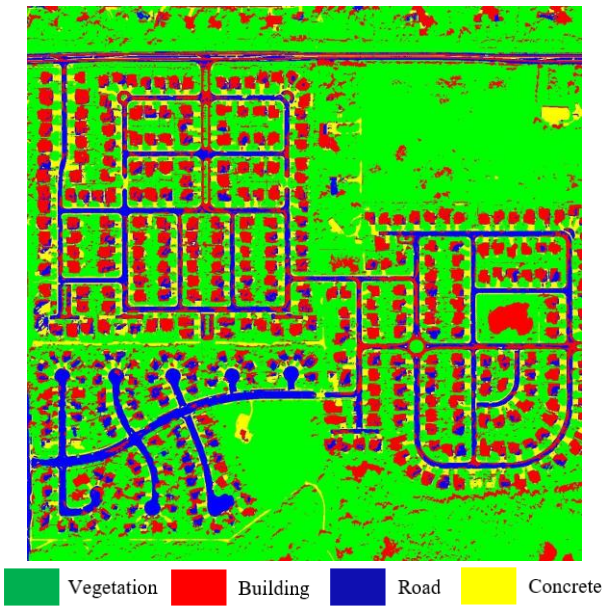

Figure 14. Decision Tree Classification

Table 7. Shows the confusion matrix of DT classifier

	Ground Truth (Percent)			
Class	Vegetation	Road	Building	Concrete
Vegetation	**68.86**	6.97	10.79	4.01
Road	7.31	**61.89**	18.27	0.62
Building	23.75	27.4	**68.12**	0.12
Concrete	0.09	3.74	2.81	**95.24**

In high resolution satellite images using the DT method has a confusion matrix as shown in Table 7. Concrete as actual class and concrete as predicted class has value confusion matrix is high when compared with other classes.

Figure 15. k-Nearest Neighbor

Table 8. The Confusion Matrix Performance of KNN classifier

	Ground Truth (Percent)			
Class	Building	Road	Vegetation	Concrete
Building	**87.37**	1.09	18.37	0.62
Road	0.24	**93.56**	8.58	0.81
Vegetation	11.12	2.92	**71.09**	2.36
Concrete	1.27	2.43	1.95	**96.2**

C. Accuracy Assessment and Comparisons

In order to assess the accuracy of classification performance, there are many metrics available in the literature. The two most popular metrics are overall accuracy (OA) and Kappa Coefficient.

The overall accuracy rate for the SVM classifier was 78.6% and 83.30% for high and low resolution satellite imagery, respectively. We found that unlike KNN classification, SVM classification was much more adept at classifying land use which is the main reason.

Table 9. Performance of accuracy classifier in percent

Image	SVM	DT	KNN
High Resolution	78.60	68.41	76.26
Low Resolution	83.30	59.08	82.34

As show in Table 9 and Table 10, with all datasets, SVM always showed the most accurate results, followed by

Decision Tree and KNN. In high resolution, the classification accuracy of SVM and DT were significantly different. However, the classification accuracy of SVM and KNN were not significantly different.

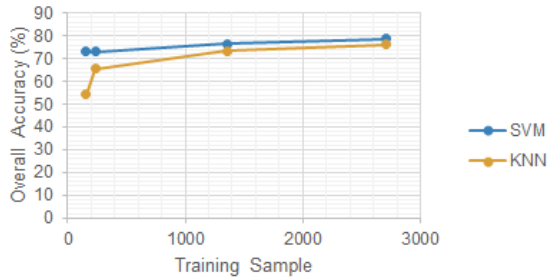

Figure 16. Training Sample comparison

Figure 16 illustrated different training sample number will affect the overall accuracy in SVM and K-NN method. It is indicated that the sample size of training samples has more impact on the classification accuracy for K-NN than for SVM. In addition, SVM has flexibility in choice of threshold than both methods.

Table 10. Performance of Kappa Coefficient

Image	SVM	DT	KNN
High Resolution	0.5967	0.4294	0.5673
Low Resolution	0.7375	0.4459	0.7253

Table 10 is performance of Kappa Coefficient. The lowest error was achieved with SVM for all datasets. Therefore, the optimal Kappa Coefficient for SVM classifier was chosen as 0.59 and 0.73 in the high and low resolution, respectively.

V. CONCLUSION

In this paper, Decision tree, SVM and k-Nearest Neighbor data mining techniques are applied to classify the region of interest from images in order to get the meaningful observations. The main objective of this paper is to help the researchers to select best technique for image classification. The SVM method has good accuracy compared to the Decision Tree and k-Nearest Neighbor methods. For High Resolution Satellite Image and Low Resolution Satellite Image has an accuracy of 78.60% and 82.34% respectively by using SVM method. It can also be seen that the value of Kappa coefficient in the SVM method has a high value compared to both methods.

ACKNOWLEDGMENT

We would like to thank Prof. Hsueh-Chan Lu, Research Instructor from Department of Geomatics, National Cheng Kung University, for creating this opportunity for research.

REFERENCES

[1] Bremner D, Demaine E, Erickson J, Iacono J, Langerman S, Morin P, Toussaint G, Output- sensitive algorithms for computing nearest-neighbor decision boundaries, *Discrete and Computational Geometry*, 2005, pp. 593 – 604.

[2] Christoper, J.C.B. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 121–167.

[3] J. Kim, B. S. Kim, and S. Savarese, Comparing Image Classification Methods: K-Nearest-Neighbor and Support-Vector-Machines, *Applied Mathematics in Electrical and Computer Engineering*, pp. 133-138.

[4] Lawrence, R., Bunn, A., Powell, S. and Zambon, M. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment. 90, 3 (2004), 331–336.

[5] Lu, D.; Weng, Q.A. Survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 2007, 28, 823–870.

[6] P. T. Noi and M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery", *Journal Sensor*, vol. 18, 2018.

[7] S. K. Dash and M. Panda, "Image Classification using Data Mining Techniques," *Advances in Computer Science and Information Technology (ACSIT)*, vol. 3, pp. 157–162, April-June 2016.

[8] Vyoma Patel, G. J. Sahani, A Survey on Image Classification using Data Mining Techniques, *IJSRD - International Journal for Scientific Research & Development*, Vol. 2, Issue 10, 2014, pp. 746 - 750

[9] Wan, S., Lei, Tc. and Chou, Ty. 2010. A novel data mining technique of analysis and classification for landslide problems. Natural hazards. 52, 1 (2010), pp. 211 – 230.