Application of support vector machine for classification of multispectral data

2014 IOP Conf. Ser.: Earth Environ. Sci. 20 012038

(http://iopscience.iop.org/1755-1315/20/1/012038)

Application of support vector machine for classification of multispectral data

Nurul Iman Saiful Bahari, Asmala Ahmad and Burhanuddin Mohd Aboobaider
Department of Industrial Computing
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

E-mail:nuruliman.sf@gmail.com, asmala@utem.edu.my and burhanuddin@utem.edu.my

Abstract. In this paper, support vector machine (SVM) is used to classify satellite remotely sensed multispectral data. The data are recorded from a Landsat-5 TM satellite with resolution of 30x30m. SVM finds the optimal separating hyperplane between classes by focusing on the training cases. The study area of Klang Valley has more than 10 land covers and classification using SVM has been done successfully without any pixel being unclassified. The training area is determined carefully by visual interpretation and with the aid of the reference map of the study area. The result obtained is then analysed for the accuracy and visual performance. Accuracy assessment is done by determination and discussion of Kappa coefficient value, overall and producer accuracy for each class (in pixels and percentage). While, visual analysis is done by comparing the classification data with the reference map. Overall the study shows that SVM is able to classify the land covers within the study area with a high accuracy.

1. Introduction

Land cover classification is one of the most important remote sensing applications. It has been widely used in many fields such as town planning, studies of environmental change, land resource planning, and geological mapping. Generally, a good classifier should be able to classify pixels into desirable land covers. The factors taken into consideration when selecting a classification method include accuracy, speed and practicality. Among the frequently used methods in classification are maximum likelihood classification (MLC) and artificial neural network (ANN). However, there are drawbacks to these classifications; ANN has been associated with over fitting and local minima problems [5], while MLC needs large training area and assumption that the data are normally distributed. In recent years, there have been an effort to develop better reliable classification methods; support vector machine (SVM) is one of them [10]. SVM is characterised by an efficient hyperplane searching technique that uses minimal training area and therefore consumes less processing time. The method is able to avoid over fitting problem and requires no assumption on data type. Although non-parametric, the method is capable of developing efficient decision boundaries and therefore can minimise misclassification. This is done through finding of optimal separating hyperplanes between classes by focusing on the training cases (support vectors) that lie at the edge of the class distributions, with the other training cases being excluded [13]. This study aims to carry out SVM classification on land covers over Selangor, Malaysia.

Originally, SVM is a binary classifier that works by identifying the optimal hyperplane and correctly divides the data points into two classes. There will be an infinite number of hypeplanes and SVM will select the hyperplane with maximum margin. The margin indicates the distance between the classifier and the training points (support vector). Figure 1 illustrates the basic idea of support vector machine [13]. A number of techniques can be used to expand the classifier from binary to multiclass i.e. one against all and

one against one [9]. SVM can classify the data linearly or nonlinearly. Kernel function is used for nonlinear data. By comparison from previous studies, SVM produces classifications with a relatively high accuracy [5], [7], [11].

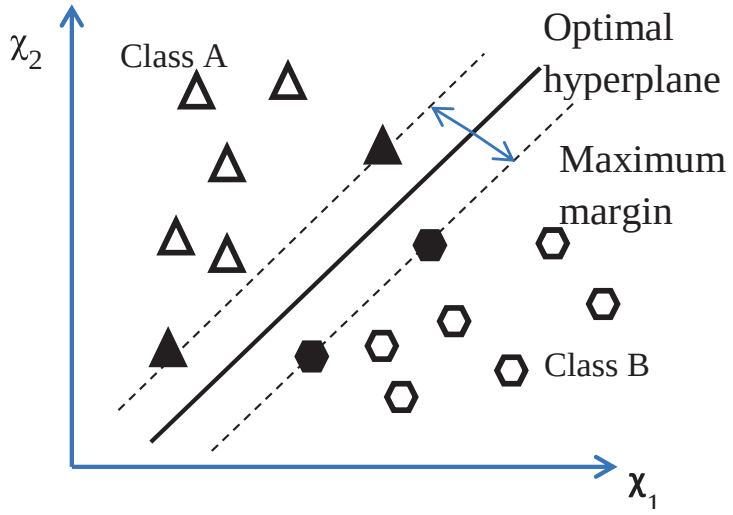


Figure 1. Basic idea of SVM.

Foody and Mathur [7] demonstrated that a single multiclass SVM classification can be applied and derive accurate classifications. They compared different classification methods, viz. discriminant analysis, decision trees, feed forward neural network, and SVM classification, and found that SVM yielded the highest accuracy. The outcome is similar to a study carried out by Shi and Yang [11] that showed SVM surpass the MLC in terms of quantitative accuracies. Candade and Dixon [5] also showed that SVM perform better than other classification. They made use of three different types of SVM kernels such as Polynomial, RBF and linear kernel and compared the results with ANN. They stated that the difference between the ANN and SVM maybe because ANN have the problem of over-fitting, local minima and sensitive to the dimensionality of the data while, SVMs show better accuracy even with a small number of training samples. In this paper, we present the application of SVM in classifying land covers recorded from the Landsat-5 TM satellite over the area of Port Kelang, located in Selangor, Malaysia.

2. Methodology

The Landsat-5 TM dataset, dated 11 Feb 1999 with a spatial resolution of 30×30m, was obtained from the Malaysia Remote Sensing Agency (ARSM). The dataset was subsetted spatially and spectrally. The area is about 840m^2 within longitude 101^{o}10'E to 101^{o}30'E and latitude 2^o 99' N to 3^{o}15'N. The Landsat-5 TM is equipped with seven bands ranging from 0.45 to 2.35 μm, however, only 6 bands were used in this study (band 1, 2,3,4,5 and 7) since band 6 is a thermal band and therefore, is not relevant to this study. Initially, the land covers were visually analysed with the aid of the reference map produced in October 1991 by the Malaysian Surveying Department and ARSM. A total of 10 land cover was identified, i.e. industrial, oil palm, rubber, coastal swamp forest, coconut, dryland forest, cleared land, bare land, and sediment plumes. Pixels that representing water, cloud and cloud shadow were masked out because this study only focuses on land area [2].

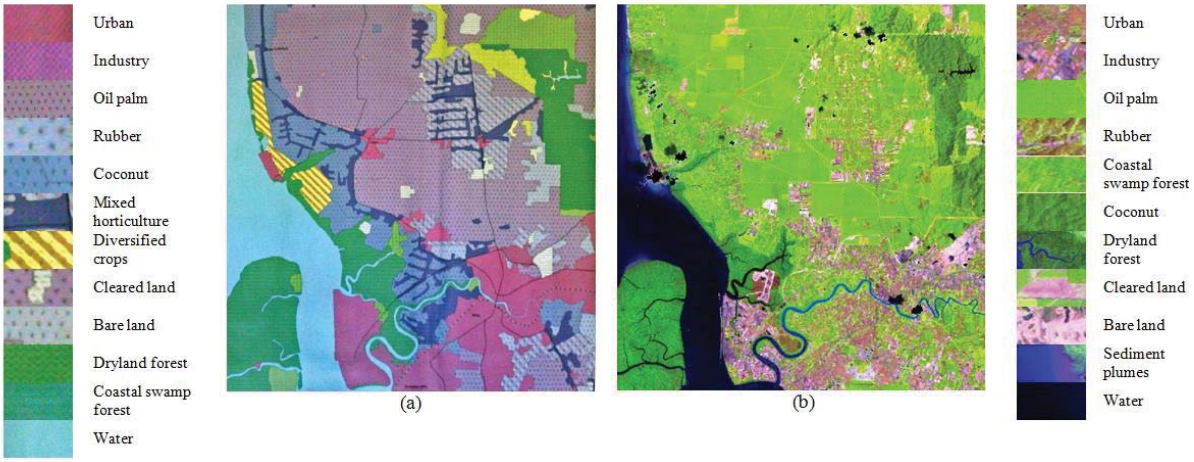


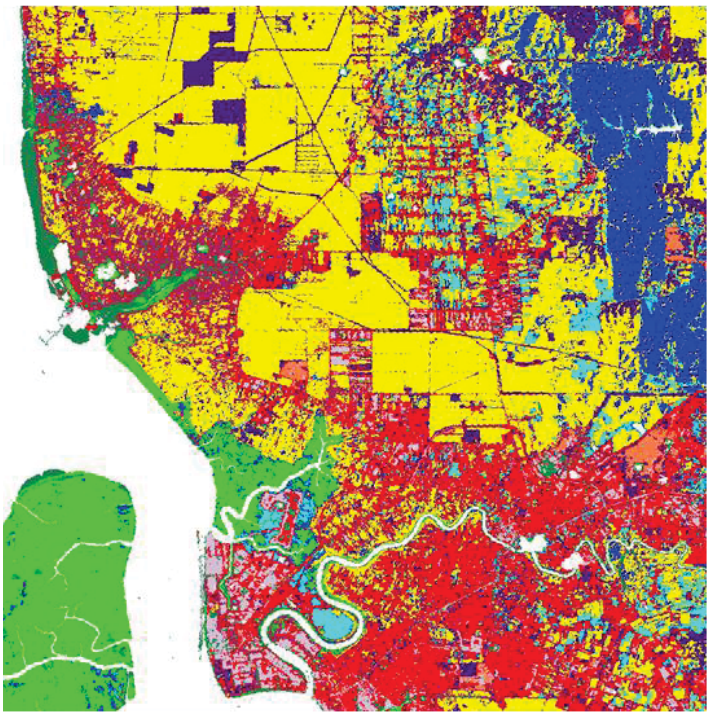
Figure 2. (a) The land cover map (b) The Landsat-5 TM with band 3, 2 and 1 in RGB layer.

Figure 2 above shows the study area at Port Klang. There are three types of plantation with palm oil as the major plantation followed by coconut and rubber. The palm oil plantation can easily recognize by the light green patches spread toward the north area. The coconut and rubber plantation are scattered at northwest and southeast part respectively. Moreover, there are two types of natural forest; in the northeast is dryland forest and coastal swamp forest in the southwest part of the study area. The industrial area (brighter pixel) covers the south part and the urban placement (pixel in red colour) is more concentrated toward southeast along the Klang River. The water bodies (pixel in blue colour) are mainly the Malacca Straits and Klang River. Sediment plume (brighter blue colour pixel) is located at the shore of the water body towards northwest part. The bare land and cleared land are scattered all over the place with the bare land more concentrated in the industrial and urban areas, while the cleared land is more concentrated at the oil palm plantation area.

Training areas were created by choosing polygons that contain training pixels representing the land covers. This step is the most crucial part, since inaccurate training pixels can lead to serious misclassification. Although SVM can classify with only small training areas, in this study, medium to large training areas are used. This is due to the fact that such training areas tend to produce classification with a high accuracy.

3. Results & discussion

Figure 3 below shows the classification performed by SVM with its legend and percentage area for each class. The 10 classes are coastal swamp forest (green), dryland forest (blue), oil palm (yellow), coconut (maroon), rubber (cyan), urban (red), industry (thistle), sediment plume (sea green), bare land (coral), cleared land (purple). Water, cloud and cloud shadow were masked white and excluded from further processing stages since this study focuses on land area only. Coastal swamp forest, dryland, oil palm, urban, industrial and sediment plume classes were easily recognised by their well-grouped pixels. While, rubber, cleared land, coconut and bare land classes are less grouped and more scattered over the places. The largest class is oil palm (34.9%), followed by urban (14.6%) and cleared land (11.9%). While two of the smallest areas classified are sediment plume (3.2%) and industrial (3.5%). Visually, overall performance of SVM in land cover classification is good as it can classify all pixels effectively.



Class	Colour	Area (%)
Coastal swamp forest		8.7
Dryland forest		7.2
Oil palm		34.9
Cleared land		11.9
Sediment plume		3.2
Coconut		10.6
Bare land		1.0
Urban		14.6
Industrial		3.5
Rubber		4.4

Figure 3. SVM classification of Landsat-5 TM data and the class area in percentage. (Water, cloud and cloud shadow are masked white).

For accuracy assessment purposes, selection of ground truth pixels was done by random sampling. Accuracy analysis was carried by comparing the classified pixels with ground truth pixels using a confusion matrix [3]. Table 1 shows confusion matrices for SVM classification in terms of (a) pixels and (b) percentage. The results were presented in terms of producer accuracy and overall accuracy. Producer accuracy is the probability that a pixel in the classification falls into class x given the ground truth class is x and can be calculated using [8]:

$$\text{Producer accuracy} = \frac{c_{aa}}{c_{\bullet a}} \times 100\% \qquad (1)$$

where,

c_{aa}= element at position a^{th} row and a^{th} column

$c_{\bullet a}$= column sums

Table 1 (c) below shows the producer accuracy for each class. Overall accuracy is the total percentage of pixels correctly classified and can be computed as:

$$\text{Overall accuracy} = \frac{\sum_{a=1}^{U} c_{aa}}{Q} \times 100\% \quad (2)$$

where,
Q = total number of pixels
U = total number of classes

The overall accuracy of 97.1% indicates that the SVM classification is good at separating pixel into their classes with very few of the pixels were not classified into its classes.

The Kappa coefficient is a measure of the agreement between variables and can be calculated using:

$$\kappa = \frac{\sum_{a=1}^{U} \frac{c_{aa}}{Q} - \sum_{a=1}^{U} \frac{c_{a\bullet} c_{\bullet a}}{Q^2}}{1 - \sum_{a=1}^{U} \frac{c_{a\bullet} c_{\bullet a}}{Q^2}} \quad (3)$$

where,
$c_{a\bullet}$ = row sums

Kappa coefficient value is always less than or equivalent to 1. A value of 1 indicates that the variables are in perfect agreement. The value of 0.96 shows that the SVM classification and the ground truth data are at a very good agreement [1]

Table 1. Confusion matrix for SVM classification.

		Ground Truth (Pixel)										
	Class	CSF	DF	OP	CL	SP	C	BL	U	I	R	Total classified pixels
Classification Data (Pixels)	CSF	3621	0	0	0	0	0	0	0	0	0	3621
	DF	0	1536	4	0	0	1	0	0	0	7	1548
	OP	0		2504	12	0	77	0	0	0	0	2593
	CL	0	0	1	209	0	2	0	26	3	2	243
	SP	0	0	0	2	445	18	0	0	5	0	470
	C	0	0	54	15	20	305	0	0	0	0	394
	BL	0	0	0	1	0	0	78	0	0	0	79
	U	0	0	1	12	0	6	0	537	0	0	556
	I	0	0	0	3	0	0	0	5	162	0	170
	R	0	0	2	4	0	0	0	0	0	165	171
	Total ground truth pixels	3621	1536	2566	258	465	409	78	568	170	174	9845

(a)

		Ground Truth (Percent)										
	Class	CSF	DF	OP	CL	SP	C	BL	U	I	R	Total classified (percent)
Classification Data (Percents)	CSF	100	0	0	0	0	0	0	0	0	0	36.78
	DF	0	100	0.16	0	0	0.24	0	0	0	4.02	15.69
	OP	0	0	97.58	4.65	0	18.83	0	0	0	0	26.37
	CL	0	0	0.04	81.01	0	0.49	0	4.58	1.76	1.15	2.47
	SP	0	0	0	0.78	95.7	4.4	0	0	2.94	0	4.77
	C	0	0	2.1	5.81	4.3	74.57	0	0	0	0	4
	BL	0	0	0	0.39	0	0	100	0	0	0	0.8
	U	0	0	0.04	4.65	0	1.47	0	94.54	0	0	5.65
	I	0	0	0	1.16	0	0	0	0.88	95.29	0	1.73
	R	0	0	0.08	1.55	0	0	0	0	0	94.83	1.74
	Total ground truth (percent)	100	100	100	100	100	100	100	100	100	100	100

(b)

Class	Producer Accuracy	
	(Pixel)	(Percent)
Coastal swamp forest (CSF)	3621/3621	100
Dryland forest (DF)	1536/1536	100
Oil palm (OP)	2504/2566	97.58
Cleared land (CL)	209/258	81.01
Sediment plume (SP)	445/465	95.7
Coconut (C)	305/409	74.57
Bare land (BL)	78/78	100
Urban (U)	537/568	94.54
Industry (I)	162/170	95.29
Rubber (R)	165/174	94.83

(c)

Almost all classes possess producer accuracy more than 90%, except for coconut and cleared land. The producer accuracy of 74.57% of coconut class is mainly because 18.83% of its pixels were classified as oil palm. This is due to the fact that coconut and oil palm has similar physical structure, producing similar spectral behaviour and misclassified as each other [1]. While for cleared land, producer accuracy of 81.01% is mainly because most of the pixels are being classified as all other classes except for coastal swamp forest and dryland forest. Cleared land representing the land without any vegetation or building like untarred road, tarred road, alley, orchard path, cleared land for next vegetation, and house divider.

However, for 30x30m resolution satellite data, all these features tend to be mixed with features from other classes like oil palm, coconut palm, residential building, industrial building, and rubber tree resulting in misclassification of the pixels. On the other hand, coastal swamp forest, dryland forest and bare land class shows 100% producer accuracy. This is because these three classes have very distinct features resulting in different spectral characteristics and not easily misclassified. In general, producer accuracy for each class shows that SVM can classify very well, except for bare land and coconut class. Such problem may be solved if higher resolution satellite data are used.

4. Conclusions

In this paper, we have studied the SVM performance for tropical land cover classification on Landsat TM data. The study area of Klang Valley has more than 10 land covers and classification using SVM has been done successfully without any pixel being unclassified. The overall accuracy and kappa coefficient of 97.1% and 0.96 respectively shows that this classifier can yield high classification accuracy and has high agreement between ground truth and classified data. It indicates that this classifier is a promising and reliable for remote sensing classification tool although not that well known compared to other methods such as ANN and MLC.

Acknowledgement
The authors thank Universiti Teknikal Malaysia Melaka (UTeM) for funding this study under FRGS grant (FRGS/2012/FTMK/TK06/03/02 F00143) and also the Malaysian Remote Sensing Agency (ARSM), Ministry of Science, Technology and Innovations (MOSTI) for providing the data.

References

[1] Ahmad A and Quegan S 2012 Analysis of maximum likelihood classification on multispectral data Applied Mathematical Sciences 6 6425 – 6436
[2] Ahmad A and Quegan S 2012 Cloud masking for remotely sensed data using spectral and principal components analysis Engineering, Technology & Applied Science Research (ETASR) 2 221 – 225
[3] Ahmad A and Quegan S 2013 Comparative analysis of supervised and unsupervised classification on multispectral data Applied Mathematical Sciences 7 74 3681 – 3694
[4] Altman DG 1991 *Practical Statistics for Medical Research* Chapman and Hall London 404
[5] Candade N and Dixon B 2004 Multispectral classification of Landsat images: A comparison of support vector machine and neural network classifiers. *ASPRS Annual Meeting Proc. Denver CO*
[6] Congalton R G and K Green 1999 *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* Lewis Publishers Boca Raton FL
[7] Foody G M and Mathur A 2004 A relative evaluation of multiclass image classification by support vector machines *IEEE Transactions on Geoscience and Remote Sensing 42* **6** 1335–1343
[8] Jensen J R 1986 *Introductory Digital Image Processing* Prentice-Hall Englewood Cliffs New Jersey 379

[9] Md Sap M N Kohram M 2008 Support Vector Classification of Remote Sensing Images Using Improved Spectral Kernels *J. Teknologi Maklumat 20*
[10] Mountrakis G, Im J and Ogole C 2011 Support vector machines in remote sensing: A review *ISPRS J. of Photogrammetry and Remote Sensing* 66 **3** 247-259
[11] Shi D and Yang X 2012 Support Vector Machine for Landscap e Mapping from Remote Sensor Imagery *AutoCarto 2012*
[12] Yao W and Han M 2011 Remote sensing image classification with parameter optimized Support Vector Machine based on evolutionary computation *Int. Workshop on Computational Intelligence (IWACI) 2011* 290–294
[13] Vapnik V 1995 *The Nature of Statistical Learning Theory* New York Springer Verlag
[14] Zhu H 2009 Classification of Urban Remote Sensing Image Based on Support Vector Machines *2009 17th Int. Conf. on Geoinformatics*