ERJ
Engineering Research Journal
Faculty of Engineering
Minoufiya University

Sensitivity of Pixel-Based Classifiers to Training Sample Size in Case of High Resolution Satellite Imagery

Doma, M. I.[1]; Gomaa, M. S.[2] and Amer, R. A.[3]

1. Civil Engineering Department, Faculty of Engineering, Menoufia University, Egypt
2. Surveying Engineering Department, Faculty of Engineering, Benha University, Egypt
3. GIS & Surveying Engineer, Menofya Governorate, Egypt

ABSTRACT

Thematic maps representing the characteristics of the Earth's surface have been widely used as a primary input in many land related studies. Classification of remotely sensed images is an effective way to produce these maps. The value of the map is clearly a function of the accuracy of the classification. Selecting proper size of samples and classification method are essential issues to produce accurate thematic maps. In the present study, training data sets at various sizes used to investigate the effect of the training set size on the classification accuracy. Six supervised classification methods with different characteristics were applied to produce land use/land cover thematic map of the study area. The used classifier include: Parallelepiped, Minimum Distance, Mahalanobis Distance, Maximum Likelihood, Neural Network and Support Vector Machine (SVM). The results showed that optimum sample size differs from classifier to another. In the case of limited number of training pixels, SVM and maximum likelihood classifiers produced higher classification accuracies than the rest of classifiers.

الخرائط المصنفة والتى تمثل خواص سطح الأرض استخدمت بتوسع كمدخل اساسى للعديد من الدراسات المتعلقة بسطح الأرض. يعتبر تصنيف صور الأقمار الصناعية طريقة فعالة لانتاج مثل هذا النوع من الخرائط. وتتوقف قيمة الخريطة المنتجة على دقة عملية التصنيف. لذا فان اختيار الحجم الملائم للعينة المستخدمة لتوجيه عملية التصنيف علاوة على عملية التصنيف نفسها هو مسألة جوهرية لانتاج خرائط مصنفة دقيقة. ***فى هذه الدراسة*** تم دراسة تأثير استخدام حجوم مختلفة من العينة اللازمة لتوجيه عملية التصنيف على الدقة الناتجة. كذلك تم استخدام ستة انواع من المصنفات ذات الخواص المختلفة لانتاج خرائط مصنفة توضح استخدامات وغطاءات سطح الأرض لمنطقة الدراسة. المصنفات المستخدمة تشتمل على: Neural - Maximum Likelihood - Mahalanobis Distance - Minimum Distance- Parallelepiped Network - Support Vector Machine (SVM) . اظهرت النتائج ان حجم العينة اللازم للحصول على افضل نتيجة يختلف من مصنف الى اخر. كما انه فى حالة استخدام حجم عينة غير كاف فان الـ SVM وكذلك الـ Maximum Likelihood اظهرت دقة تصنيف اعلى مقارنة بباقى المصنفات.

1. INTRODUCTION

Data from satellite sensors has become an important tool for researchers studying land use and land cover change. Remote sensing offers the advantage of rapid data acquisition of land use information at a lower cost than ground survey methods and the analysis of this data can provide critical insights into the evolving human environment relationship. Thematic maps derived from remotely sensed data are used in many applications, including as input parameters to models, as source of regionally extensive environmental data, or as basis of policy analysis. Meaningful and consistent measures of thematic map reliability are necessary for the map user to assess the appropriateness of the map data for a particular application; additionally, the accuracy of the thematic map may significantly affect the outcome of an application. Measures of map accuracy are equally important for the producer of a thematic map to analyze sources of error and weaknesses of a particular classification strategy.

Measures of map accuracy are well established in the literature (e.g., Story and Congalton, 1986; Congalton, 1991; Stehman, 1997; Congalton and Green, 1999). Most commonly, accuracy assessment involves the comparison of a classified thematic map with the classification of randomly selected samples of reference data (Stehman, 1997). The most widely used measures of accuracy are derived from an error matrix (Congalton, 1991; Foody, 2002). It is worth mentioning that no one classification will be optimal from the viewpoint of each different user (Lark, 1995; Brown et al., 1999).

Training samples primarily collected on a per-pixel basis to reduce redundancy and spatial-autocorrelation. They selected through image interpretation with intensive field visits over this area. Although more training samples are usually beneficial, as they tend to be more representative to the class population, a small number of training samples is obviously attractive for logistic reasons (Li et al, 2014). It is often recommended that a

training sample size for each class should not be fewer than 10–30 times the number of bands (Van Niel et al. 2005).

Sample size is an important consideration when assessing the accuracy of remotely sensed data. Each sample point collected is expensive and therefore sample size must be kept to a minimum and yet it is critical to maintain a large enough sample size so that any analysis performed is statistically valid (Congalton, 1991).

The number of samples for each category can also be adjusted based on the relative importance of that category within the objectives of the mapping or by the inherent variability within each of the categories.

Supervised classification is the process of using training data for assigning class labels to unknown pixels. It has been widely used in remote sensing area. It is generally underlined that there is a strong relationship between classification accuracy and training data sets used in the learning stage of supervised classification method (Zhuang et al., 1994; Foody, 1999; Pal and Foody, 2010). Foody and Mathur (2006) indicated that the accuracy of a supervised image classification is a function of the training data used. With many classification algorithms, no previous study has reported an optimal number of training samples, to test the sensitivity of an algorithm to the size of training samples (Congcong et al., 2014).

After describing the study areas and data sources in the following section, this paper is organized as follows. Section 3 describes the methods. Section 4 presents and evaluates the results and the results are summarized in Section 5.

2. Study Area and Data Sources

The study area chosen for this research covers approximately 520*270m, and located in the western part of Luxor city, Egypt. A pan-sharpened QuickBird satellite image covering the study area was used to determine land cover and land use types, see figure 1. Additionally, field surveys were applied using a handheld GPS to collect ground reference information. After the detailed analysis of ground reference data, it was decided that mainly six land use and land cover types covers the study area, which are: buildings; green areas; roads; ground; water; and platforms as shown in figure 2.

Class “ground” mainly corresponds to grass, parking lots and bare fields. All recognizable features independent of their size were digitized. Adjacent buildings that were joined but obviously separated were digitized as individual buildings; otherwise, they were merged into one polygon. Larger areas covered by trees were digitized as one polygon.

Figure (1): The Quick Bird satellite image of the study area.

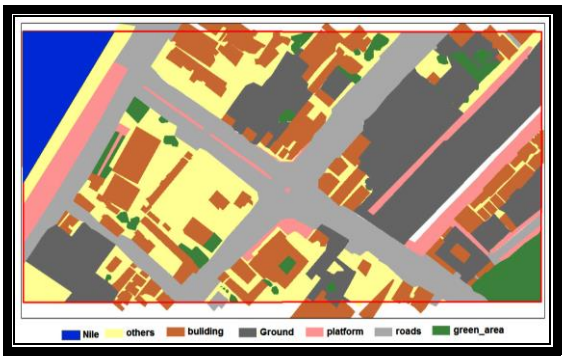

Figure (2): The truth image.

3. Methodology

In this study two investigations have been conducted, the first one uses equal sizes of training data (100, 200, 300, and 400) pixels of six classes (building, green area, road, ground, water and platforms) for the classification of the QuickBird image. The Second experiment uses the previous recommendation of Congalton (1991), which recommended to concentrate the sampling on the categories (classes) of interest and increase their number of samples while reducing the number of samples taken in the less important categories. With small addition which is making a Proportion and fit the sample size for each class and its area (number of pixels) in the image depending on the optimum ratio of training size used in Kavzoglu and Colkesen (2012), 0.42%. With the note of all parameters of the used classifiers are set to the default of the software (ENVI 4.4), the sample size is the only one that has been changed through the two investigations. The experimental work was implemented in several stages as shown if figure 3:

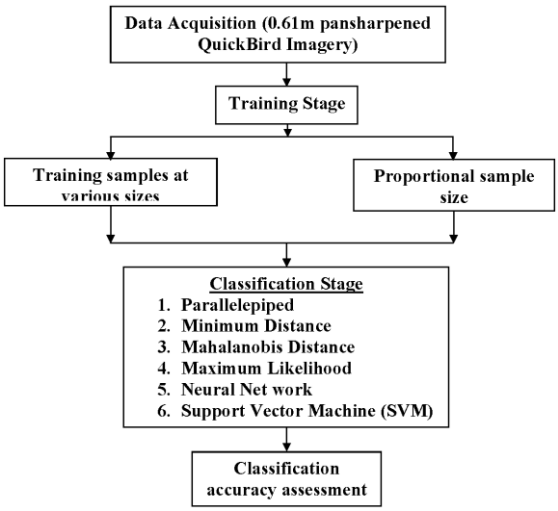

Figure (3): Procedure of the proposed comparison.

Figure 4 shows the confusion matrix and some common measures of classification accuracy that may be derived from it. The highlighted elements represent the main diagonal of the matrix that contains the cases where the class labels depicted in the image classification and ground data set agree, whereas the off-diagonal elements contain those cases where there is a disagreement in the labels.

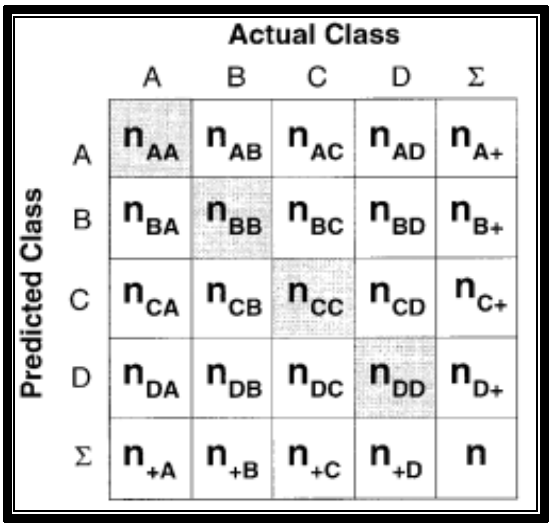

Figure (4): The confusion matrix.

$$\textbf{Percentage correct} = \frac{\sum_{k=1}^{q} n_{kk}}{n} \times 100 \quad \text{..... (1)}$$

$$\textbf{User's accuracy} = \frac{n_{ii}}{n_{i+}} \quad \text{..... (2)}$$

$$\textbf{Producer's accuracy} = \frac{n_{ii}}{n_{+i}} \quad \text{..... (3)}$$

The producer's accuracy relates to the probability that a reference sample (photo-interpreted land cover class in this project) will be correctly mapped and measures the errors of omission (1 - producer's accuracy). In contrast, the user's accuracy indicates the probability that a sample from land cover map actually matches what it is from the reference data (photo-interpreted land cover class in this project) and measures the error of commission (1- use's accuracy).

4. Results and Analysis

4.1 Using training samples at various sizes

Table 1 and figure 5 show that SVM and Mahalanobis Distance performed the best with overall accuracy of 80% for both, followed by Maximum Likelihood with 78.33% overall accuracy. In terms of training sample size, the following results have been obtained:

1. The best performing training sample size for SVM is between 200 and 300 pixels per class.
2. The best performing training sample size for both Neural Network and Parallelepiped is around 100 pixels per class.
3. The best performing training sample size for Mahalanobis Distance is around 200 pixels per class.
4. The best performing training sample size for Maximum Likelihood is between 100 and 200 pixels per class.
5. The best performing training size sample for Minimum Distance is around 300 pixels per class.

Previous results show that optimum training sample size differs from classifier to another. All classifiers are shared in the same behavior of after critical point (optimum training size sample), the classification accuracy showed downward trend; it was negatively affected with the increasing number of training pixels. Moreover, it is clear that in the case of limited number of training pixels, SVM and Maximum Likelihood classifiers produced higher classification accuracies than the rest of classifiers. It is worth mentioning that for SVM, highest overall accuracy of 80% was achieved with the training data set containing totally 1200 pixels (200 pixels per class). These results conform to (Kavzoglu and Colkesen, 2012) who explained that

considering training set size, classification performances of SVM improved till a certain level.

The overall accuracy has relatively stabilized values with the increasing of the sample size for Minimum Distance, Mahalanobis Distance, Maximum Likelihood and SVM. On the other hand, the value of the overall accuracy has greatly fluctuated with the sample size increasing for Parallelepiped and Neural Network classifiers.

Table (1): Results of different training sizes. The highlighted cell shows the maximum overall accuracy for each classifier.

Classifier	Overall accuracy %				
Parallelepiped	43.33	31.67	30	30	38.33
Min. Distance	63.33	66.67	70	66.67	58.33
Mahalanobis Distance	70	80	78.33	78.33	71.67
Max. Likelihood	78.33	78.33	71.67	70	71.67
Neural Network	55	48.33	48.33	33.33	48.33
SVMs	78.33	80	80	75	76.67
Sample size /class	100	200	300	400	Prop.
Total sample size	600	1200	1800	2400	1667

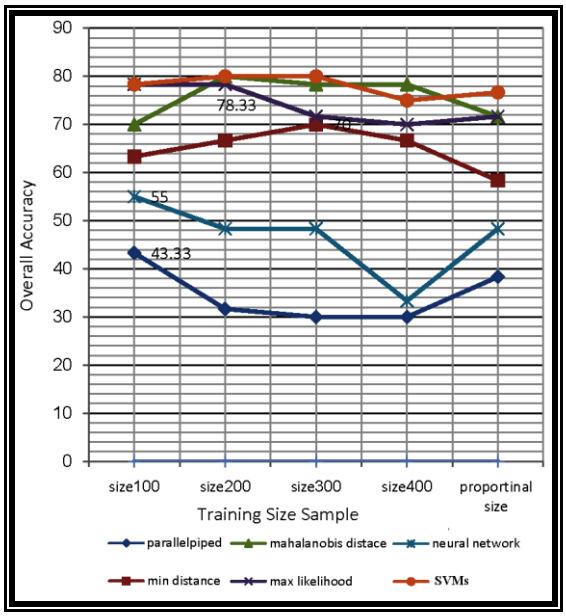

Figure (5): Overall accuracy for each classification type for different training sizes sample.

4.2 Using Proportional sample size

According to Congalton (1991) stratified random sampling is recommended where a minimum number of samples are selected from each category. Sometimes it is better to concentrate the sampling on the categories of interest and increase their number of samples while reducing the number of samples taken in the less important categories. Also it may be useful to take fewer samples in categories that show little variability such as water and increase the sampling in the categories that are more variable such as urban areas.

In this study a procedure has been applied to calculate a proportional training sample size depending on the optimum ratio used in Kavzoglu and Colkesen, (2012) which is 0.42%. First, a trust image of the study area digitized using GIS software. After that, using the software, the number of pixels for each class has been determined, and then the percentage of each class calculated. Finally, the required sample size was calculated. As compared to Kavzoglu and Colkesen, (2012), a sample of 10500 pixels for 1735*1442 QuicBird image, a sample of 1616 pixel is required for the 865*445 QuicBird image used in our experiments. This sample distributed to each class as its percentage as shown in Table 2.

Table (2): Size of proportional sample per class. B: Buildings; GA: Green Areas; R: Roads; G: Ground; W: Water; P: Platforms.

Class	No. of pixel	Percent %	Calculated training size (in pixels)	Actual training size (in pixels)
B	67447	20.9	338	334
GA	29861	9.3	150	155
R	83117	25.8	417	421
G	94580	29.4	475	489
W	19203	6	97	126
P	27764	8.6	139	137
Total	321972	100	1616	1662

Figure 6 shows the overall accuracy obtained for each classifier in the case of proportional training size sample. The results clearly indicates that SVMs still performing the best in terms of overall accuracy. On the other hand, the overall accuracies of all classifiers have been reduced as compared with the optimum previous results. The reduction in overall classification accuracy caused by each classifier as compared with case one (equal sizes of training samples) was determined and summarized in table 3. Whereas the application of Min. Distance classifier resulted in a maximum reduction of 11.67%, the application of SVM resulted in a

minimum reduction of 3.33%. The most important notice is that, there was no decrease in the overall accuracy for the Neural Network. However, all classifiers have agreed that with the proportional sample size, the overall accuracy is comparable to the best overall accuracy. On the other hand, many contradictions are notable in the performance of the classifiers with different training sample sizes. One possible reason for the discrepancies in the results of overall accuracy can be the convergence of the platform layer in the value of the spectral resolution with the building layer which, would lead to significant misclassifications between the two layers of buildings and platform and this is confirmed by visual inspection of the results as shown in figure 7. And also confirmed by the results of confusion matrix as the producer and user accuracy for the platform layer are 30% and 60 % respectively. Therefore, there is a necessity to rely on other data and information than the values of the spectral resolution such as spatial and/or spectral attributes.

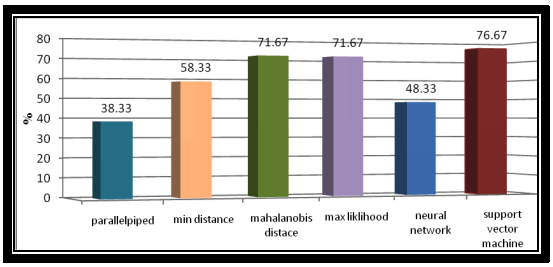

Figure (6): The overall accuracy obtained for each classifier in the case of proportional training size sample.

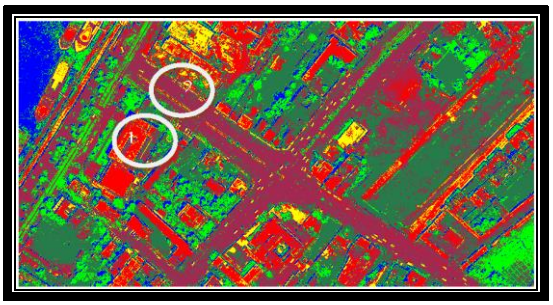

Figure (7): Classification result of SVM. White circles show misclassification between building (1) and platform (2).

Table (3): The reduction in overall classification accuracy caused by each classifier as compared with case one, equal sizes of training samples.

Classifier	Reduction %
Parallelepiped	5
Min. Distance	11.67
Mahalanobis Distance	8.33
Max. Likelihood	6.66
Neural Network	0
SVM	3.33

5. Conclusions

In this study, Five training sample sizes (100, 200, 300, 400 and proportional size) were compared for six supervised classifiers which include: Parallelepiped, Minimum Distance, Mahalanobis Distance, Maximum Likelihood, Neural Network and Support Vector Machine (SVM). The classifiers tested on a pixel-based classification. All supervised classifiers tested with five sets of different sized training samples. The results showed that optimum sample size, which gives the highest overall accuracy, differs from classifier to another. All classifiers are shared in the same behavior of after critical point (optimum training size sample), the classification accuracy showed downward trend. Moreover, in the case of limited number of training pixels, SVM and Maximum Likelihood classifiers produced higher classification accuracies than the rest of classifiers. The convergence of one class in the value of the spectral resolution with another class leads to significant misclassifications. Therefore, it is necessary to rely on other data and information than the values of the spectral resolution such as spectral and/or spatial attributes.

6. References

Brown, J. F., Loveland, T. R., Ohlen, D. O., & Zhu, Z. 1999. The global land-cover characteristics database: the user's perspective. Photogrammetric Engineering and Remote Sensing, 65, 1069 – 1074.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35–46, 1991.

Congalton, R.G. and Green, K. 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and practices (Boca Raton, London, New York: Lewis Publishers).

Congcong Li, Jie Wang, Lei Wang, Luanyun Hu and Peng Gong. 2014. Comparison of C classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat

Thematic Mapper Imagery Remote Sens. 2014, 6, 964-983; doi:10.3390/rs6020964.

Dicks, S. E., & Lo, T. H. C. 1990. Evaluation of thematic map accuracy in a land-use and land-cover mapping program. Photogrammetric Engineering and Remote Sensing, 56, 1247 – 1252.

Foody, G.M. 2002. Status of land covers classification accuracy assessment. Remote Sensing of Environment, vol. 80, pp. 185-201.

Foody, G.M., Mathur, A., Sanchez-Hernandez, C. and Boyd, D.S. 2006. Training set size requirements for the classification of a specific class. Remote Sensing of Environment, 104, 1–14.

Kavzoglu, T and Colkesen, I. 2012. The effects of training set size for performance of support vector machines and decision trees, International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences Florianopolis-SC, Brazil, July 10-13, 2012.

Lark, R. M. 1995. Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. International Journal of Remote Sensing, 16, 1461 – 1480.

Li, C.; Wang, J.; Wang, L., Hu, L. and Gong, P. 2014. Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery. Remote Sens. 2014, 6, 964-983; doi: 10.3390/rs6020964.

Pal, M., Foody, G.M. 2010. "Feature selection for classification of hyperspectral data by SVM". IEEE Transactions on Geoscience and Remote Sensing, Vol. 48(5): 2297-2307.

Stehman, S.V. 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment, 62, 77 – 89. Powell, R. L., Matzke, N., de Souza, C., Clark, M., Numata, I., Hess, L. L., et al. (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sensing of Environment, 90, 221−234.

Story, M. and Congalton R.G. 1986. Accuracy assessment: A user's perspective. Photogrammetric Engineering and Remote Sensing 52(3): 397-399.

Van Niel, T., McVicar, T., Datt, B. 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. Remote Sensing of Environment, 98, 468–480.

Zhuang, X., B.A., Engel, Lozanogarcia D.F., Fernandez R.N., Johannsen C.J. 1994. Optimization of training data required for neuro - classification. International Journal of Remote Sensing, Vol. 15(16): 3271-3277.