Journal of Geomatics Vol 9 No. 1 April 2015

Sensitivity of pixel-based classifiers to training sample size in case of high resolution satellite imagery

M.L. Doma[1], M.S. Gomaa[2] and R.A. Amer[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
Email: mohamed.doma@sh-eng.menofia.edu.eg

(Received: Jul 01, 2014; in final form: Jan 10, 2015)

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 a function of the accuracy of the classification. Selection of proper size of samples and classification method are important factors which govern accuracy of thematic maps. In the present study, training data sets of various sizes are used to investigate their effects on the classification accuracy. Two investigations have been carried out. The first one makes use of equal sizes of training data for the classification of 0.6 meter spatial resolution QuickBird-2 satellite image. The second experiment allocates higher sampling fraction for the classes of interest while reducing the number of samples in the less important categories. Six supervised classification methods with different characteristics are applied to produce land use/land cover thematic map of the study area. The classifiers used in the study include: Parallelepiped, Minimum distance, Mahalanobis distance, Maximum likelihood, Neural Net work and Support Vector Machine (SVM). After certain fraction of sample size, the classification accuracy showed downward trend with the increasing number of training pixels. In the case of limited number of training pixels, SVM and maximum likelihood classifiers showed higher classification accuracies than the rest of classifiers. In the case of proportional training size sample, the overall accuracies of all classifiers have been reduced as compared with the first experiment except for SVM classifier.

Keywords: Classifiers, Training sample, High resolution satellite imagery

1. Introduction

In the past few years, data from satellite sensors has become an important tool for researchers studying land use and land cover (LULC) 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.

Land cover classifications are widely used processes in the field of remote sensing. The general aim of land cover classification is the associations of each pixel within the image with a specific land cover class to produce precise classification maps of the area. 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 and as basis of policy analysis (Waske, 2007).

A variety of approaches for image classification is available and can be divided into two major groups, unsupervised and supervised. In unsupervised classification, also called clustering, only some parameters are required to be specified by the user. These parameters usually include number of clusters, criteria for convergence and no. of iterations. Then a clustering method uses these parameters to uncover statistical patterns that are inherent in the data. Examples of conventional clustering methods are ISODATA and k-means algorithm (Richards and Jia, 2003; Duda et al., 2000). It should be noted that the statistical patterns identified are just clusters of pixels with similar spectral characteristics. They do not necessarily correspond to any meaningful characteristics of ground objects. Consequently, after unsupervised classification the user must attach the actual meaning to the resulting classes since the algorithm does not provide any final membership decision (Jensen, 1996). A general problem of unsupervised algorithms is that data can consist of clusters with different shapes and sizes. Furthermore, an applicable definition of clusters and the selection of an adequate indicator for similarity are difficult (Jain et al., 2000).

Supervised classification requires apriori knowledge about the image data such as which types of land-use exist in the study area and spatial locations of representative samples for each type. The procedure of supervised classification is more controlled by the user than unsupervised classification.

Training samples for each land use type are first collected by ground surveys, or from aerial photos, maps or visual interpretation. Different collection strategies, such as single pixel, seed, and polygon, may be used (Chen and Stow, 2002). Signatures, which are statistically based criteria for each class, are then generated from the training samples. Finally the pixels

Journal of Geomatics Vol 9 No. 1 April 2015

in the image are sorted into classes according to the signatures, by use of a classification decision rule (Lillesand and Kiefer, 2004).

Training samples primarily collected on a per-pixel basis are used to reduce redundancy and spatial-autocorrelation. They are 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. It is important to maintain 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. 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 et al. (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).

Despite the fact that results of object-oriented classification of high resolution images are of considerable interest to researchers, segmentation is not an easy task and requires adjusting the segmentation parameters according to the situation. This makes the object-oriented classification process iterative and manual. Per-pixel classification is a preferred method because satellite data sets are acquired digitally on the basis of pixel units (Garcia-Gutierreza et al., 2009). The objective of this paper is to investigate the sensitivity of pixel-based classifiers to training sample size in case of high resolution satellite imagery

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 0.6 meter spatial resolution and pan-sharpened image over the area of study were collected in July, 2010 by QuickBird-2 satellite (QB02) and supplied in a TIFF digital format. The image is supplied in a product level LV2A and product type standard. This image is radiometrically adjusted to improve the radiometric quality (see figure 1). Table 1 summarises the spatial and spectral characteristics of the image used (DigtalGlobe, 2006).

Table 1: Characteristics of the image used

Spatial resolution	0.6 m
Spectral resolution	blue (450-520 nm) green (520-600 nm) red (630-690 nm)

Additionally, field surveys were carried out using a handheld GPS to collect ground reference information. After the detailed analysis of ground reference data, it was decided that mainly six primary classes of interest covers the study area, which are: built-up areas; 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. Buildings that were connected were digitized as individual buildings. Larger areas covered by trees were digitized as one polygon.

Figure 1: The Quick bird satellite image of the study area

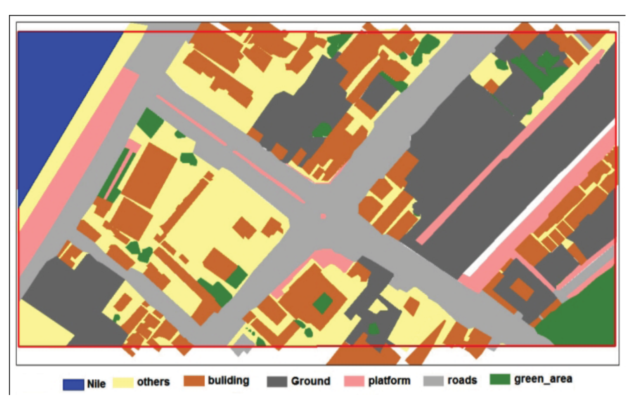

Figure 2: The truth image

3. Methodology

3.1 Image classification

In this study two investigations have been carried out. The first study makes use of equal sizes of training data (100, 200, 300, and 400) pixels of six classes (building, green areas, road, ground, water and platforms) for classification of the QuickBird image. The second experiment implements the recommendation of Congalton (1991) 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. A small addition makes 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) i.e. 0.42%. The experimental work was implemented in several stages as shown if figure 3.

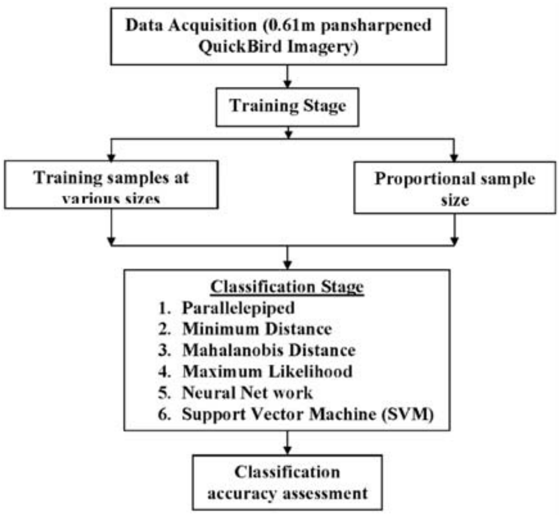

Figure 3: Procedure of the proposed comparison

3.2 Accuracy assessment

Meaningful and consistent reliability measures of thematic map reliability are necessary for the map user to access the appropriateness of the map data for a particular application. 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; 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 (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).

The overall accuracy for each of the classifications was assessed using the reference data and based on equation 1:

$$OCA = \frac{NCP}{NRP} \qquad (1)$$

where OCA is the overall classification accuracy; NCP is the total number of correctly classified pixels and NRP is the total number of reference pixels.

4. Results and analysis

4.1 Using training samples at various sizes

Table 2 and figure 4 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 both SVM and Neural Net work is between 200 and 300 pixels per class.
2. The best performing training sample size for Mahalanobis distance is around 200 pixels per class.
3. The best performing training sample size for Maximum likelihood is between 100 and 200 pixels per class.
4. The best performing training size sample for Minimum distance is around 300 pixels per class.
5. The best performing training size sample for Parallelepiped is around 100 pixels per class.

Above results show that optimum training sample size varies from one 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 observation of Kavzoglu and Colkesen, (2012) who explained that considering training set size, classification performances of SVMs 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 SVMs. 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 2: Results of different training sizes with platform class. The highlighted cell shows the maximum overall accuracy for each classifier

Classifier	Overall accuracy %				
Parallelepiped	43.33	31.67	30	30	38.33
Minimum distance	63.33	66.67	70	66.67	58.33
Mahalanobis distance	70	80	78.33	78.33	71.67
Maximum 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

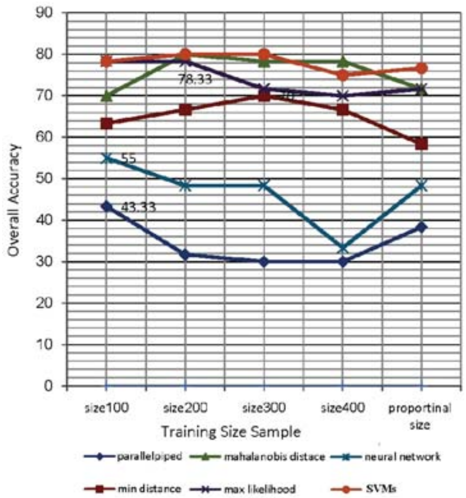

Figure 4: Overall accuracy for each classification type with platform class for different training sizes sample

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 ArcGIS 9.1 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 experiment. This sample distributed to each class as its percentage as shown in Table 3.

Table 3: Size of proportional sample per class

Class	No. of pixel	Percent %	Calculated training size (in pixels)	Actual training size (in pixel)
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

B: Buildings; GA: Green Areas; R: Roads; G: Ground; W: Water; P: Platforms

Figure 5 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 4. Whereas the application of Minimum distance classifier resulted in a maximum reduction of 11.67%, the application of SVMs resulted in a minimum reduction of 3.33%. The most important observation is that there was no decrease in the overall accuracy for the Neural Network. However, all classifiers have shown that with the proportional sample size, the overall accuracy is comparable with the best overall accuracy. On the other hand, many contradictions are noticed 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 6. 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.

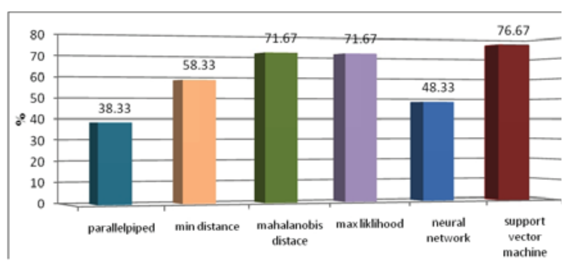

Figure 5: the overall accuracy obtained for each classifier in the case of proportional training size sample

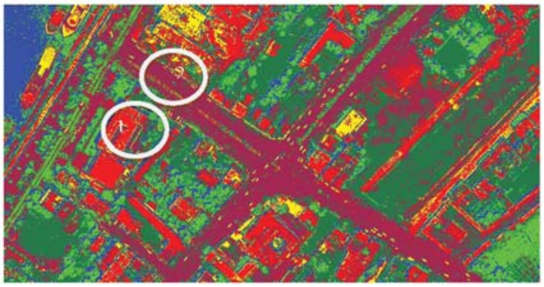

Figure 6: Classification result of SVM. White circles show misclassification between building (1) and platform (2)

Table 4: The reduction in overall classification accuracy caused by each classifier as compared with case one, equal sizes of training samples

Classifier	Reduction %
Parallelepiped	5
Minimum distance	11.67
Mahalanobis distance	8.33
Maximum likelihood	6.66
Neural Network	0
SVMs	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 Net work and Support Vector Machine(SVM). The classifiers have been tested on a pixel-based classification. This investigation has shown that optimum sample size depends on. All classifiers shared in the same behaviour 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. The convergence of one class in the value of the spectral resolution with another class leads to significant misclassifications. Therefore, spectral and/or spatial information from satellite imagery can also be applied along with image data in order to improve the performance of the classification process and to refine the results. Based on the results obtained in this research, future work is required to use medium resolution (5m or 25m) data for image classification. This would be more appropriate to bring out the importance of sample size versus methodology.

References

Brown, J., T. Loveland, D. Ohlen and Z. Zhu (1999). The global land-cover characteristics database: The user’s perspective. Photogrammetric Engineering and Remote Sensing, 65, pp. 1069 – 1074.

Chen, D. and D. Stow (2002). The effect of training strategies on supervised classification at different spatial resolution. Photogrammetric Engineering and Remote Sensing, 68, pp. 1155–1162.

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

Congalton R. and K. Green (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. 6, pp. 964-983; doi: 10.3390/rs6020964.

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

DigtalGlobe, (2006). QuickBird Imagery Products – Product Guide. Revision 4.7.1

Duda, R., P. Hart and D. Stork (2000). Pattern classification. 2nd Ed., Chichester, New York, John Wiley & Sons.

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

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

Garcia-Gutierreza, J., L. Goncalves-Secob and J. Riquelme- Santosa (2009). Decision trees on LIDAR to classify land uses and covers. In F. Bretar, M. Pierrot-Deseiligny and G. Vosselman (Eds.). Proceedings of the ISPRS workshop on Laser scanning-XXXVIII /W8, 1-2 September 2009, Paris, France.

Jain, A., R. Duin and J. Mao (2000). Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), pp. 4-37.

Jensen, J. (1996). Introductory digital image processing: A remote sensing perspective. 2nd Ed., Prentice Hall, Upper Saddle River.

Kavzoglu, T. and I. Colkesen (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. (1995). Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. International Journal of Remote Sensing, 16, pp. 1461 – 1480.

Li, C., J. Wang, L. Wang, L. Hu and P. Gong (2014). Comparison of classification algorithms and training sample sizes in urban Land Classification with Landsat Thematic Mapper Imagery. Remote Sensing, 6, pp. 964-983; doi: 10.3390/rs6020964.

Lillesand, T. and R. Kiefer (2004). Remote sensing and image interpretation. Fourth Edition, John Willey & Sons, Inc., New York.

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

Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment, 62, pp. 77 – 89.

Powell, R., N. Matzke, C. de Souza, M. Clark, I. Numata, L. Hess (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sensing of Environment, 90, pp. 221−234.

Richards, J. and X. Jia (2003). Remote sensing digital iImage analysis: An introduction. 3rd ed., Springer, New York.

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

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

Waske, B. (2007). Classifying multisensor remote sensing data: Concepts, algorithms and applications. PhD thesis, Bonn University, Germany.

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