A graphical approach for the evaluation of land-cover classification procedures

PENG GONG and PHILIP J. HOWARTH

Earth-Observations Laboratory, Institute for Space and Terrestrial Science, Department of Geography, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

(Received 8 June 1989; in final form 6 September 1989)

Abstract. A method for evaluating the effectiveness of different feature combinations and training strategies is described. Preliminary tests have been made using two groups of feature combinations derived from SPOT High Resolution Visible (HRV) data and two sets of training samples. The method is objective, and needs no ground confirmation or interaction from the image analyst. It is recommended as a surrogate for detailed accuracy assessment when attempting to find an optimum set of training pixels or feature combinations for image classification.

1. Introduction

Accuracy estimates of land-use and land-cover classifications serve two groups of people, the users and the analysts. For the users, accuracy estimates indicate the reliability of the classification results. For the analysts, however, they are indicators of the effectiveness of the classification procedures used in the study. The classification procedures may involve different feature combinations, training strategies or classifiers. It is the analyst's task to identify the optimum classification procedures to give the highest classification accuracy.

An ideal accuracy assessment method requires ground information at the time of data acquisition (Richards 1986). However, it is often difficult to collect such field data due to the remoteness of the study area or the time constraints of the project. These make the use of existing data essential. In addition, it has been demonstrated that obtaining accurate ground information is a difficult task (Curran and Williamson 1985). As a result, when selecting reference data to make accuracy assessments, researchers often have to use information interpreted from aerial photographs or ground data which were collected at another time (Campbell 1987). Accuracy estimates obtained in this manner will inevitably be biased by errors of subjectivity emanating from the photointerpretation or errors caused by changes in the landscape between dates of acquisition of the imagery and the ground information. These estimates may in turn affect the analyst's decision when selecting an appropriate classification procedure. Furthermore, experience has shown that it is time-consuming to conduct such accuracy assessments.

Several well-known methods are routinely used to improve image classification, such as the transformed-divergence approach in feature selection (Richards 1986) and use of scatter diagrams or equiprobability contours when examining training quality (Lillesand and Kiefer 1987). However, as they are based solely upon training samples, these methods are not directly related to the final classification accuracy which is estimated from randomly selected test samples. Decisions on selecting the most

0143–1161/90 $3.00

appropriate feature combinations or training sets on this basis are inappropriate if one wishes to obtain the optimum classification result.

It is desirable, therefore, for the analyst to have available a simple testing procedure which is based on random test samples, is objective and can be applied before the classification algorithm is run. In this Letter, a method which satisfies these criteria is described. The procedure involves examination of the average logarithmic probabilities for a group of pixels to help the analyst select appropriate feature combinations and training strategies for multispectral classification.

2. Method

The method is based on the assumption of a multivariate normal model for each class of land cover or land use. As indicated by Richards (1986), this assumption is widely adopted in remote sensing multispectral classification using algorithms such as the maximum-likelihood classifier. For illustration, we assume that the data to be classified are one-dimensional. Figure 1 shows two possible pixel values (X_1 and X_2). The three normal curves are assumed to be the probability distributions $p(C_i)$, ($i=1$, 2, 3) for three classes (C_1, C_2, C_3). The probabilities of pixel X_1 being classified into the three classes are $P_1(C_1)$, $P_1(C_2)$ and $P_1(C_3)$. Similarly for pixel X_2 the probabilities are $P_2(C_1)$, $P_2(C_2)$ and $P_2(C_3)$. As can be seen in figure 1, pixel X_2 has less ambiguity than pixel X_1 in terms of being classified into class C_2. From the analyst's point of view, more pixels with fewer ambiguities (such as pixel X_2) are preferred. This is also the major purpose in most classification research in which attempts are made to select and develop training strategies, feature combinations and classifiers to reduce ambiguities in pixel labelling. Such ambiguities are reflected by the differences among a pixel's probabilities of being labelled in each class in a descending order, which is referred to as the 'order of significance'. The two largest probabilities are particularly important, because the closer the second probability is to the first, the more likely it is that ambiguities will be created by the classification decision. As

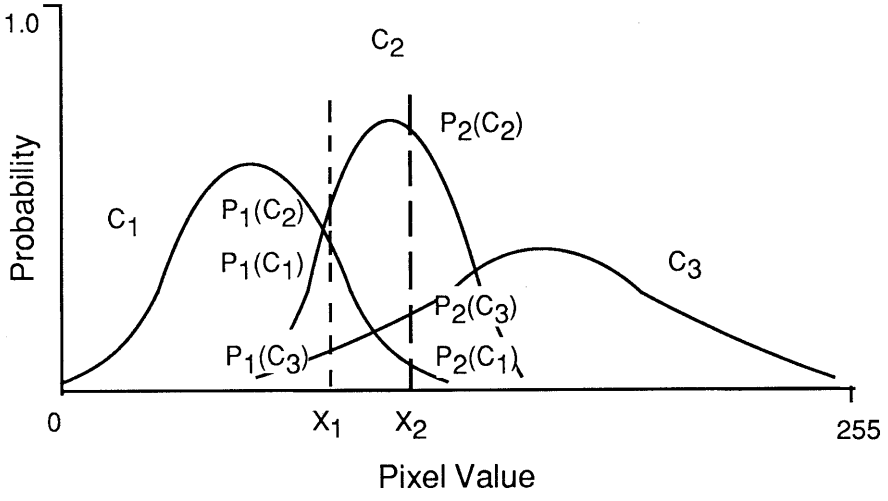

Figure 1. Illustration of classification ambiguities for two selected pixels, X_1 and X_2. Further explanation is given in the text.

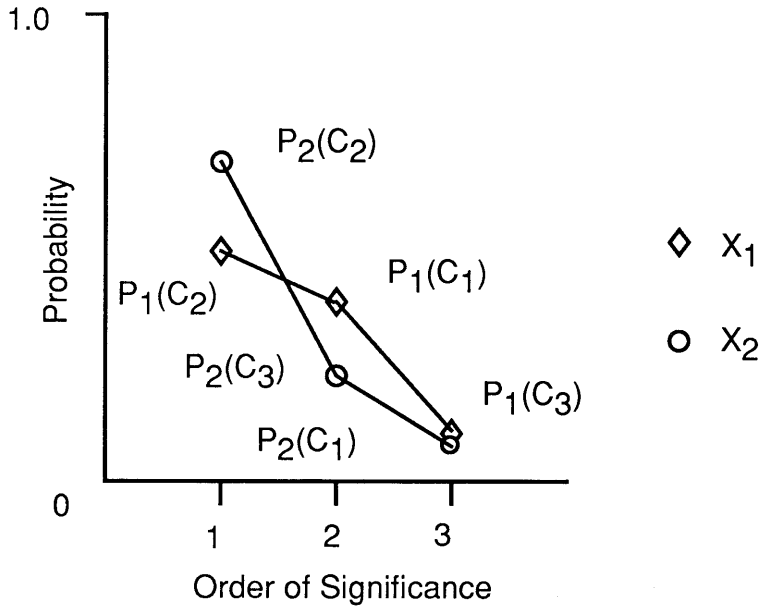

Figure 2. Probabilities in the order of significance for the two selected pixels.

shown in figure 2, the overall differences for pixel X_2 are larger than those for pixel X_1. Looked at another way, the general slope of the curve representing the ordered probabilities for pixel X_2 is steeper than that for pixel X_1. This means that selection of the most appropriate training pixels can be determined by identifying the training samples which produce the least number of ambiguous pixels (i.e. pixels similar to X_2 rather than X_1). This is equivalent to examining the magnitudes of the average probability differences or the average slopes; the larger the magnitude, the better are the training pixels. The same logic can be applied when evaluating the most appropriate feature combinations to use in image classification.

To obtain a general trend for the probability curves for a group of pixels, the average probability at each order of significance is obtained. The average probabilities created in this way are then plotted as a graph. The curve which is produced is referred to as the 'probability trend curve'. It is used to evaluate the training capabilities of different selections of pixels and the appropriateness of different feature combinations. For a given number of m training statistics (m classes) and a given group of N sample pixels, we can calculate the probabilities for each pixel X according to the Bayesian rule

$$P(C_i|X)=P(C_i)(2\pi)^{-n/2}|\Sigma_i|^{-1/2}\exp(\tfrac{1}{2}(X-M_i)^{\mathrm{T}}\Sigma_i^{-1}(X-M_i)/P(X) \qquad (1)$$

where, C_i indicates one among the m classes, n is the number of image bands, $P(C_i)$ is the *a priori* probability for class C_i, $P(X)$ is the probability for X in the image data, $|\Sigma_i|$ is the determinant of the variance-covariance matrix of class i, Σ_i^{-1} is the inverse of the variance-covariance matrix for class i, M_i is the mean of class i, and the other variables in the equation are identified in figures 1 and 2. For computational simplification, the natural logarithm is applied to the probabilities. This results in a

logarithmic probability which takes the form

$$\ln P(C_i|X) = \ln P(C_i) - 0{\cdot}5n \ln (2\pi) - 0{\cdot}5 \ln |\Sigma_i| - 0{\cdot}5(X - M_i)^{\mathrm{T}} \Sigma_i^{-1} (X - M_i) - \ln P(X) \quad (2)$$

In this case, equation (2) becomes the exponential part of the probability of pixel X having a class i. In general, the *a priori* probability for $P(C_i)$ is unknown, so that the $P(C_i)$s are assumed to be equal. The second term on the right-hand side of equation (2) is a constant. For each specific pixel X, $P(X)$ also does not change. The unchanged parts of the equation do not contribute to further analysis and, therefore, equation (2) can be further simplified as

$$\ln P(C_i|X) = -0{\cdot}5 \ln |\Sigma_i| - 0{\cdot}5(X - M_i)^{\mathrm{T}} \Sigma_i^{-1} (X - M_i) \quad (3)$$

where, $\ln P(Ci|X)$ represents the logarithmic probability value for pixel X to be classified into the ith class. The m logarithmic probabilities obtained from equation (3) for each pixel are then ranked into their order of significance. The average of the logarithmic probability values with the same order are calculated for all sample pixels. The average for each order is then plotted onto a graph for evaluation.

3. Experiment

The data used for this study are SPOT High Resolution Visible (HRV) multispectral imagery at a 20 m × 20 m spatial resolution. They were obtained on 4 June 1987 over the Town of Markham, which is situated at the rural-urban fringe of north-eastern Toronto, Canada. Because there exists a correlation of 0·96 between the two visible bands of the image, a principal component transformation was applied to the original three bands. To observe the effectiveness of the method in feature selection, two groups of feature combinations were tested. The first feature group consisted of the two principal component (PC) images containing over 98 per cent of the total variance. In the second feature group, a structural information (SI) band consisting of an edge-density image created from the original HRV Band 1 image (Gong and Howarth 1990) was used in combination with the two PC bands. In addition, two different training strategies were tested to see the effectiveness of the method in evaluating training quality. The training strategies were single-pixel training and block training. The two different feature groups and the two different training strategies constitute four feature-training groups based upon which four classifications were produced using the maximum-likelihood classifier. For each classification, the same land-cover classification scheme and the same reference data were used for accuracy estimation.

Land-cover types in the classification scheme were residential roof, paved surface, industrial and commercial roof, cleared land, lawn and tree complex, cultivated grass, deciduous tree, coniferous tree, mature crop, new crop and pasture, bare field and water surface. These land-cover types were defined based on their spectral identities which make them suitable for application of a per-pixel classifier, and also on the ease with which information classes (land-use classes) can be derived from the land-cover types. A group of 1024 test pixels identified by the stratified systematic unaligned sampling strategy (Jensen 1983) was obtained from the test images. These pixels were used to generate both the probability trend curves and the accuracy estimates.

To generate accuracy estimates, the test pixels were identified on the 1:8000 scale aerial photographs. These aerial photographs were recorded nearly 2 months before the date of the satellite overpass. Field studies were also undertaken in June 1988. The

Table 1. Overall accuracies obtained from applying the maximum-likelihood classifier with different training strategies to different feature combinations.

Training strategy	Feature combination	Overall accuracy (per cent)
Single pixel	PC images	46·8
Single pixel	PC and SI images	56.3
Block	PC images	40·2
Block	PC and SI images	46·4

results of the pixel identifications were then used as reference data for comparison with the results of the maximum-likelihood classifications. The overall accuracy was used as a summary index to represent the accuracy for each confusion matrix (table 1). The accuracy estimates indicate that for a given feature group single-pixel training shows higher accuracies than block training. Also, for a given training set, inclusion of the SI band in a feature group improves the classification accuracy. Tests show that all these improvements are significant at the 95 per cent confidence level.

Following the preparations described above, a test of the method was made. A probability trend curve was generated for each of the four feature-training combinations by applying the method described in § 2 to the selected test pixels. These curves were then compared with the accuracy estimates shown in table 1 to determine the effectiveness of the method.

4. Results and discussion

In table 2, the logarithmic probability values which were calculated with the proposed method are listed. The probability trend curves for the logarithmic probability values of the top four orders of significance are shown in figure 3. As can be seen, the shapes of the probability trend curves are as anticipated. For a given feature combination, the two curves obtained with single-pixel training are steeper, therefore indicating a higher accuracy, than the curves generated using block training.

Table 2. The ordered average logarithmic probabilities.

Order of significance	Training strategies and feature combinations			
	Two PC images		Two PC images and SI image	
	Single	Block	Single	Block
1	−4·53	−4·43	−7·22	−7·22
2	−5·92	−5·77	−10·30	−9·50
3	−7·91	−7·25	−13·47	−12·21
4	−11·89	−8·72	−17·69	−15·82
5	−18·05	−12·30	−23·67	−20·64
6	−22·25	−16·66	−28·52	−27·10
7	−28·92	−20·50	−35·62	−33·38
8	−36·06	−26·54	−45·93	−41·22
9	−49·82	−33·88	−56·83	−47·80
10	−67·69	−49·66	−73·66	−68·92
11	−88·13	−79·64	−94·53	−87·28
12	−99·83	−91·39	−100·00	−96·19

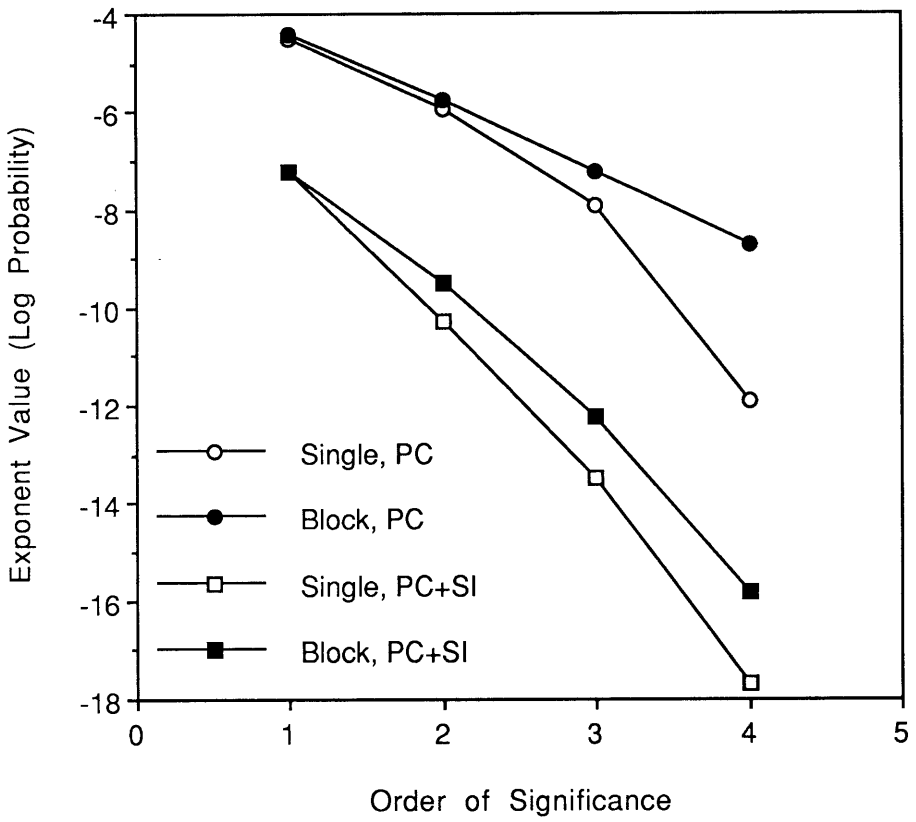

Figure 3. The probability trend curves showing the top four orders of significance recorded in table 2.

For a given training strategy, the curve obtained with the combined PC and SI bands is steeper than the one produced using the PC image alone. This means that when only one classification procedure is changed (such as the training strategy or the feature combination), the steepness of the probability trend curves can show the analyst which is the better choice. In other words, when the analyst is only concerned about making the best choice among several groups of training strategies or several different feature combinations, the probability trend curve can qualitatively point out the most appropriate one to use.

It can be seen from figure 3 that, no matter what kind of training strategy has been used, the curves obtained using the PC and SI images combined are steeper than those produced using the two PC images. It thus appears that the change of feature combination is more significant than the change of training strategy. When table 1 is checked, however, this is not the case. Thus, the graphical approach outlined in this paper would seem to be effective only for changes being compared within each individual classification procedure, such as within different training strategies or within feature selection. When mixed classification procedures are compared, this graphical evaluation method may not supply the correct answer.

By examining the difference, or the slope between the first two orders of significance on each curve, similar results to those discussed in the above paragraph

are obtained. This supports the suggestion made in § 2 that it is the first two significant values that are most important in the assessment. The difference, derived by subtracting the logarithmic probability value of the second order from the first in each curve (as indicated in table 2), may be used as the simplest single index to represent a probability trend curve.

5. Conclusions

From the above results, it would appear that the method proposed for the evaluation of classification procedures can be used to qualitatively evaluate different training strategies and the effectiveness of different feature combinations. This is provided that only the training strategies or the feature combinations are compared at one time. The method is objective and reduces the time-consuming task of accuracy checking. What is perhaps needed in the long term is a more sophisticated index which is capable of evaluating the whole probability trend curve.

Acknowledgments

The authors gratefully acknowledge the assistance of SPOT Image Corporation of France and the Canada Centre for Remote Sensing in supplying the SPOT HRV data used in this study as part of the Programme d'Évaluation Préliminaire SPOT (PEPS), Project No. 229. This research is funded by a Centre of Excellence grant from the Province of Ontario to the Institute for Space and Terrestrial Science and NSERC Operating Grant A0766 awarded to P. J. Howarth. Mr. Gong's studies are supported by the International Development Research Centre (IDRC) of Ottawa. The authors would like to thank David Barber for his comments on a draft of this paper.

References

CAMPBELL, J. B., 1987, *Introduction to Remote Sensing* (New York, London: The Guilford Press).

CURRAN, P. J., and WILLIAMSON, H. D., 1985, The accuracy of ground data used in remote-sensing investigations. *International Journal of Remote Sensing*, **6**, 1637–1651.

GONG, P., and HOWARTH, P. J., 1990, The use of structural information for improving land-cover classification accuracies at the rural-urban fringe. *Programmetric Engineering and Remote Sensing*, **56**, 67–73.

JENSEN, J. R., 1983, Urban/suburban land use analysis. In *Manual of Remote Sensing*, 2nd edition, edited by R. N. Colwell (Falls Church: American Society of Photogrammetry), pp. 1571–1666.

LILLESAND, T. M., and KIEFER, R. W., 1987, *Remote Sensing and Image Interpretation*, 2nd edition (New York: John Wiley & Sons).

RICHARDS, J. A., 1986, *Remote Sensing Digital Image Analysis: An Introduction* (Berlin, Heidelberg, New York, London, Paris, Tokyo: Springer-Verlag).