A Modified Contextual Classification Technique for Remote Sensing Data

K.M.S. Sharma and A. Sarkar

Abstract

Conventional techniques of classification make use of the spectral information at each pixel to predict the class of that pixel independently of the observations at other pixels. Contextual techniques, on the other hand, utilize the information from other neighboring pixels also. Two methods of contextual classification exist — one for low-resolution data and one for high-resolution data. A new method is proposed by combining these two methods. The new method is compared with the Gaussian maximum-likelihood classification and the two methods of contextual classification for low-resolution and high-resolution data. Classification algorithms are compared using normalized classification accuracies and the Kappa statistics.

Introduction

An important objective of remote sensing is land-cover classification. Each object/land-cover class may have its own characteristic spectral response in different spectral bands of the electromagnetic spectrum (Swain and Davis, 1978). This characteristic feature of the land-cover classes is helpful in their identification and delineation in a multispectral image. In the absence of noise, the spectral response of two pixels from the same land-cover class would be identical in feature space (response at different spectral bands forming the features). In practice, however, the presence of noise causes the spectral response of a particular land-cover class to deviate from its ideal response. This noise is usually of two types, class dependent noise and class independent noise. Class dependent noise is caused by the inherent natural variations present in a particular land-cover class, while class independent noise is due to the other sources of variation such as scatter and/or absorption, deterioration of the sensors, etc. The objective of classification in remote sensing is, therefore, to partition the noisy image into its constituent classes.

In the usual approach, the spectral response vectors of each class are modeled to have multivariate normal distributions, and the parameters of such models are estimated from training samples. Such a technique is known as supervised classification. In an unsupervised classification, they are estimated from test image pixels by a suitable clustering method. The pixel class assignments are based on likelihood calculated from the observations of each pixel to belong to each of the classes under consideration. In this technique, known as the maximum-likelihood classification technique, the class of a pixel at a location (i,j) is decided based solely on the observations at that pixel. Thus, the decision for the classification of a pixel is made independently of other pixels. Even though such a procedure is used extensively and with a fair amount of success, such a procedure is likely to lead to misclassifications in the presence of random noise. The presence of random noise causes different classes to resemble each other. To overcome such a problem, a class of techniques called contextual techniques have evolved which make use of the spatial context of a pixel in its classification. These techniques are based on the assumption that the response and class of two spatially neighboring pixels are highly related. For example, if (i,j) and (k,l) are two neighboring pixels and if (i,j) belongs to class k, then there is a high probability that (k,l) also belongs to same class. Thus, the decision for a pixel is to be taken based not only on the observation x_{ij} at (i,j) but also on all x_{kl} where (k,l) is a neighbor of (i,j). Use of contextual techniques will usually result in a reduction of classification error rates (Swain *et al.*, 1981; Jhung and Swain, 1996; Welch and Salter, 1971).

Several contextual classification techniques have been developed for the classification of remote sensing data. A brief review of some of these techniques is given in the next section. The proposed method is then presented. Methods of accuracy analysis are given, the test results are presented, and, finally, a summary and conclusions are given.

Review of Existing Methods

The different approaches adopted by the researchers during the past few years to incorporate context in the classification of remotely sensed data can be grouped as follows:

- Methods based on the classification of homogeneous objects,
- Techniques based on probabilistic relaxation,
- Methods derived using compound decision theory and sequential compound decision theory, and
- Methods derived based on a stochastic model for the distribution of classes in the scene.

One of the early and best known contextual classifiers is Supervised Extraction and Classification of Homogeneous Objects (SECHO), discussed by Landgrebe (1980). It can be regarded as a "per field" classifier. It first divides the scene into homogeneous image segments and then classifies these segments using an extended version of the Gaussian maximum-likelihood (GML) algorithm. This technique is well known to be efficient for classifying data sets that contain homogeneous objects that are large compared to the spatial resolution provided by the sensor. The CASCADE algorithm of Merickel *et al.* (1984) and the agglomerative clustering technique of Amadsun and King (1988) are other examples of such an approach.

Contextual techniques based on relaxation methods iteratively adjust some initial estimates of class membership probabilities by reference to spatial context. The possibility of using probabilistic relaxation for contextual classification of remotely sensed data has been investigated by various au-

Department of Mathematics, Indian Institute of Technology, Kharaggpur-721 302, India (anjan@maths.iitkgp.ernet.in).

Photogrammetric Engineering & Remote Sensing, Vol. 64, No. 4, April 1998, pp. 273–280.

0099-1112/98/6404–273$3.00/0

thors (Eklundh *et al.*, 1980; Peleg, 1980; Richards *et al.*, 1982). DiZenzo *et al.* (1987) proposed an efficient implementation of the probabilistic relaxation method suited to the needs of actual remote sensing applications. The probabilistic relaxation scheme proposed by DiZenzo *et al.* (1987) works as follows. Let $i=1, 2, ..., N$ be the N pixels to be classified into K classes $\omega_1, \omega_2, ..., \omega_K$. Let the vector $[P_i(\omega_1), P_i(\omega_2), ..., P_i(\omega_K)]$ denote the probabilities of pixel i belonging to classes $\omega_1, \omega_2, ..., \omega_K$, $0 \le P_i(\omega_l) \le 1 \sum_{l=1}^{K} P_i(\omega_l) = 1$. The classes are assumed to be mutually exclusive and exhaustive. For each pair of neighboring pixels i and j and each pair of classes ω_l and ω_k, there is a compatibility measure $r_{ij}(\omega_k, \omega_l)$ in the range $[-1, 1]$ or $[0, 1]$. These coefficients indicate the degree to which classifying pixel i to ω_k and pixel j to ω_l are compatible. The iterative updating process uses the updating rule defined by

$$p_i^{(t+1)}(\omega_k) = \frac{p_i^{(t)}(\omega_k)\,(1 + q_i^{(t)}(\omega_k))}{\sum_{l=1}^{K} p_i^{(t)}(\omega_l)\,(1 + q_i^{(t)}(\omega_l))}$$

where

$$q_i^{(t)}(\omega_k) = \frac{1}{N_b} \sum_{j=1}^{N_b} \sum_{l=1}^{K} r_{ij}(\omega_k, \omega_l)\, p_j^{(t)}(\omega_l).$$

N_b is the number of neighbors considered for pixel i. The r_{ij} can be chosen empirically. One possibility is to use estimates of the mutual information of the pair of events $i\epsilon\omega_k$, $j\epsilon\omega_l$ as $r_{ij}(\omega_k, \omega_l)$. Emperically, it can be estimated from the GML classification as

$$r_{ij}(\omega_k, \omega_l) = \frac{N_{ij}(\omega_k, \omega_l)}{\sum_{k=1}^{K} N_i(\omega_k, \omega_l) \sum_{l=1}^{K} N_j(\omega_k, \omega_l)}$$

where $N_{i,j}(\omega_k, \omega_l)$ is the frequency of occurrence of classes ω_k and ω_l as neighbors at pixel i and j. It is important that, in order to obtain reliable estimates of $r_{i,j}$, the GML classification must be reasonably accurate. If this requirement is not met, the coefficients $r_{i,j}$ will not reflect the actual characteristics of the image. In that case, they cannot be expected to promote true improvements of classification accuracy.

The process of probabilistic relaxation does not utilize measurement information except in the initialization stage when the measurement information is used to obtain the initial class membership probabilities for each pixel.

Let the image data consist of a two-dimensional array of $N=N_1 \times N_2$ random observations $X_{i,j}$ having fixed but unknown classifications $\theta_{i,j}$. The observation $X_{i,j}$ consists of p-dimensional measurements while $\theta_{i,j}$ can be any one of the K spectral or informational classes from the set $G=\{\omega_1, \omega_2, ..., \omega_k\}$.

Let $\mathbf{X}$ denote a vector whose components are the ordered observation $\mathbf{X}=[X_{i,j} \mid i=1,2, ..., N_1, j=1,2, ..., N_2]^T$; similarly, let $\mathbf{\Theta}$ be the vector of states (true classifications) associated with the observations in $\mathbf{X}$: $\mathbf{\Theta}=[\theta_{i,j} \mid i=1,2,...,N_1, j=1,2,...,N_2]^T$.

Let the action (classification) taken with respect to pixel (i,j) be $a_{i,j}$ and the loss suffered due to such action be $\lambda(\theta_{i,j}, a_{i,j}(X))$. The average loss over all classifications is then $\frac{1}{N}\sum_{i,j} \lambda(\theta_{i,j}, a_{i,j}(X))$. The expected average loss or risk is $R_\theta = E[\frac{1}{N}\sum_{i,j} \lambda(\theta_{i,j}, a_{i,j}(X))]$ where E denotes the expectation operator. The action $a_{i,j}$ is in general a function of all observations X. The corresponding decision rule which minimizes the risk R_θ assigns pixel (i,j) to class ω_k if

$$P(\theta_{i,j} = \omega_k \mid X) \ge P(\theta_{i,j} = \omega_l \mid X),\ \forall l = 1,2,...,K. \quad (1)$$

This decision rule is based on the context of all observations X. Swain *et al.* (1981) derived a decision rule based on a subset of observations in X which includes X_{ij}, the observation at pixel (i,j) to be classified. The subset denoted by D_{ij} constituted the neighborhood of X_{ij}. They assume $a_{i,j}=d(D_{i,j})$ where $d(.)$ is a decision rule. It is shown that the risk R_θ is minimized if $d(D_{i,j})$ is the action a which minimizes

$$R_\theta = \sum_{\theta^p} G(\theta^p)\, \lambda(\theta^p, a)\, f(D_{i,j} \mid \theta^p)$$

where θ^p is the vector of classes corresponding to the elements of the set $D_{i,j}$, and $G(\theta^p)$ is the context distribution, i.e., the relative frequency with which θ^p occurs in the array $\mathbf{\Theta}$, with the class corresponding to pixel (i,j) being a. They further assume that the observations are class conditionally independent so that $f(D_{i,j} \mid \theta^p) = \Pi_{k\epsilon N_{i,j}} f(X_l \mid \theta_l)$. The parameters of the class conditional distribution $f(X_l \mid \theta_l)$ are obtained from the training samples from each class. To estimate the context distribution $G(\theta^p)$, they use a non-contextual preclassified image. They then propose an iterative classification scheme where, after each classification, the context distribution is re-estimated and the samples are re-classified. This process is continued until no more changes in classification occur for any pixel. The performance of this method has been shown for various neighborhood sizes using simulated data. But their performance on real data was far from satisfactory.

Application of the decision rule above (Equation 1) using Bayes' procedure depends upon the specification of the probability distribution of $(\theta_1, \theta_2, ..., \theta_N, x_1, ..., x_N)$. This being an enormous vector, Equation 1 will usually be impossible to use in practice. An approximation to it is to use a smaller set of observations $\Delta_{i,j}$ instead of X as was done by Hjort and Mohn (1984).

Hjort *et al.* (1985) expressed the feature vectors from neighboring pixels as a sum of two independent processes, one having independent vectors with class dependent distributions and the other being a contaminating autocorrelated noise process. This model is combined with different models for the joint distribution of classes of the neighbors, giving contextual classification rules incorporating the spatial relationship of the feature vectors. Accordingly, the decision rule (Equation 1) is modified to classify pixel (i,j) to class k if

$$P(\theta_{i,j} = \omega_k \mid \Delta_{i,j}) \ge P(\theta_{i,j} = \omega_l \mid \Delta_{i,j}),\ \forall l = 1,2,...,K \quad (2)$$

where $\Delta_{i,j}$ is the set of observations from a chosen neighborhood of pixel (i,j). Thus, the rule (Equation 2) is contextual in the sense that observations from the neighbors are also used in classifying the center pixel. Assuming a nearest-neighbor system with $\Delta_{i,j} = (x_{i,j}, x_{i-1,j}, x_{i,j+1}, x_{i+1}j, x_{ij-1})$, the *a posteriori* probability can be expressed as

$$P(\theta_{i,j} = \omega_k \mid \Delta_{i,j}) = \frac{1}{f(\Delta_{i,j})} P(\omega_k) \sum_{a,b,c,d} g(a,b,c,d \mid \omega_k)\, f(\Delta_{i,j} \mid \omega_k, a,b,c,d) \quad (3)$$

where (a,b,c,d) is one of the K^4 possible class configurations for the neighbor pixels and $g(a,b,c,d \mid \omega_k)$ is the conditional probability of seeing this configuration given that the central pixel is from class ω_k. Further, $f(\Delta_{i,j} \mid \omega_k, a,b,c,d)$ is the conditional joint density of the five vectors given that the classes in question are ω_k, a,b,c,d. The denominator $f(\Delta_{i,j})$ is the unconditional density of the five spectral vectors, making $\Sigma_k P(\theta_{i,j}=\omega_k \mid \Delta_{i,j}) = 1$. They have proposed a simple autocorrelation model for modeling the class independent spatial correlation of observations to specify the joint distribution of neighboring pixel vectors. For this purpose, they assumed

that the observed process $X_{i,j}$ can be decomposed into two independent components as

$$X_{i,j} = Y_{i,j} + E_{i,j}$$

where $Y_{i,j}$ is a purely class dependent process having no spatial correlation and $E_{i,j}$ is a class independent zero mean process having a spatial correlation structure. The class conditional distribution of Y is assumed to be multivariate normal for each class, having a common covariance matrix; that is, $Y_{i,j} \mid \theta_{i,j} \sim N_p(\mu_{\theta_{i,j}}, (1-\theta)\Sigma)$ and $\text{cov}(Y_{i,j}, Y_{k,l}) = 0 \ \forall \ (i,j) \neq (k,l)$. The process $E_{i,j}$ is specified by $E_{i,j} \sim N_p(0, \theta\Sigma)$ and $\text{cov}(E_{i,j}, E_{k,l}) = \rho^d \theta\Sigma$, where $d^2 = (i-k)^2 + (j-l)^2$.

The classification rule is obtained by an expression for $f(\Delta_{i,j} \mid \omega_k, a,b,c,d)$ in Equation 2. Under the Gaussian assumption, this is also a multivariate normal density in $5p$ dimensions. Assuming that the conditional distribution of x_{ij}, given the classes in the neighborhood, just depends on the class of pixel (i,j), the class density $f(x_{i,j} \mid \omega_k)$ can be factored out. This gives

$$P(\theta_{i,j} = \omega_k \mid \Delta_{i,j}) = \text{const.}\ P(\omega_k)\, f(x_{i,j} \mid w_k)\, R_k(x_{i-1,j}, x_{i,j+1}, x_{i+1,j}, x_{i,j-1}).$$

$R_k(x_{i-1,j}, x_{i,j+1}, x_{i+1,j}, x_{i,j-1})$ can be considered as an adjustment factor accounting for the contextual contribution in the classification rule.

$$R_k(x_{i-1,j}, x_{i,j+1}, x_{i+1,j}, x_{i,j-1}) = \sum_{a,b,c,d} g(a,b,c,d \mid \omega_k) h(x_{i-1,j}, x_{i,j+1}, x_{i+1,j}, x_{i,j-1} \mid x_i, \omega_k, a,b,c,d).$$

Depending upon the models assumed for g and h, different contextual classification rules result. The functions g and h contribute to the contexual information. The g-function is used to model the distribution of classes in the chosen neighborhood, and its contribution is in the form of probabilities of different class configurations. The h-function is used to model the joint distribution of the pixel intensities in the chosen neighborhood given their classes.

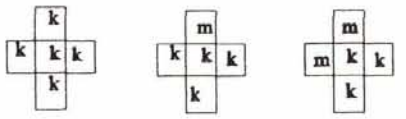
A popular assumption in the literature (Swain *et al.*, 1981; Hjort *et al.*, 1985; etc.) on contextual methods is that the spectral vectors are conditionally independent given the classes. But this assumption may be acceptable when applied to low resolution scenes, whereas models reflecting dependence between $X_{i,j}$ given the classes is needed in the case of high-resolution scenes. Different specifications for $g(a,b,c,d \mid \omega_k)$ are possible. A simple choice would be to assume $g(a,b,c,d \mid \omega_k) = P(a)\, P(b)\, P(c)\, P(d)$. This results in a classification rule which ignores spatial dependence between classes, but takes into account spatial dependence between spectral measurements. Another important assumption would be to consider the distribution of classes as a realization of a stochastic process. With this assumption, $g(a,b,c,d \mid \omega_k) = P(a \mid \omega_k)\, P(b \mid \omega_k)\, P(c \mid \omega_k)\, P(d \mid \omega_k)$, where $P(m \mid t) = P(\theta_{i,j} = m \mid \theta_{k,l} = t)$, (i,j) and (k,l) are immediate neighbors. The assumption provides a possible method for incorporating transition probabilities in the classification procedure. If it is further assumed that spectral vectors are conditionally independent, the classification rule reduces to

$$P(\theta_{i,j} = \omega_k \mid \Delta_{i,j}) = \text{const.}\ P(\omega_k)\, f(X_{i,j} \mid \omega_k) T_k(x_{i-1,j}) T_k(x_{i,j+1})\, T_k(x_{i+1,j})\, T_k(x_{i,j-1}) \quad (4)$$

where $T_k(x) = \Sigma_m P(m \mid k) f(x \mid m)$. This rule was derived by Welch and Salter (1971) and Haslett (1985). Swain *et al.* (1981) and Haralick and Joo (1986) also assumed class conditional independence between the spectral measurements in the neighborhood.

Methods for estimating the parameters of the conditional joint density function, assuming multivariate normal distribution for the observational vectors, are also given by Hjort *et al.* (1985). These parameters include the spatial correlation parameters in addition to the mean vector and the covariance matrix of each class.

In order to implement this contextual classification, it is necessary to estimate the configuration probabilities $g(a,b,c,d \mid \omega_k)$. They assumed positive probability for three patterns, which are referred as the X, T, and L types as shown below:



This is a realistic assumption when the regions are large compared to pixel sizes so that most first-order neighbors are formed from only two classes. There are four possible L-type patterns and four possible T-type patterns, and P(X) + P(L) + P(T) = 1.

Kartikeyan *et al.* (1994) proposed two methods, one for low-resolution data and one for high-resolution data on the lines of Khazenie and Crawford (1990). They assume spatial dependence for high-resolution data and class conditional independence for low-resolution data. The suitability of their method for classification is demonstrated with two sets of remotely sensed data of high and low resolution.

Proposed Method

Having reviewed some of the important contextual classification techniques used in remote sensing, we now consider the methods proposed by Kartikeyan *et al.* (1994). They have developed two methods, one for high-resolution data and one for low-resolution data. The terms high and low resolution refer to the value of the ratio of region size to pixel size. Even though the pixel size is fixed within an image, the region size may vary for different types of land-cover classes, and, hence, the assumption of high/low resolution uniformly for all classes in the data may not be proper in most of the cases. Thus, further investigations are needed. One possibility is to use an appropriate model for each class, depending on whether it is of high resolution or low resolution, which has been tried in this work.

The low and high resolution correspond to the value of the ratio of region size to pixel size for each class. Region size for a class is the average minimum size of a set of contiguous pixels of that class in the image. Because the pixel size in a given image is fixed, high or low resolution correspond to the region size of a class.

In the proposed method, the decision regarding high or low resolution is made based upon the actual observed configuration probabilities from a noncontextual preclassified image. If the probability of observing X,T,L type context for a class is close to unity, then for that class the data are to be considered as high resolution and the spatial correlation model is to be used for class conditional density of neighboring pixel vectors. Otherwise, class conditional independence assumption will be used and the class conditional joint density will be estimated as the product of marginal densities. In other words, the decision function is calculated as

$$g(k \mid \Delta_{i,j}) = \theta g_1(k \mid \Delta_{i,j}) + (1 - \theta)\, g_2(k \mid \Delta_{i,j}) \quad (5)$$

where

$$g_1(k \mid \Delta_{i,j}) = P(\omega_k) \sum_{a,b,c,d} P(\Delta_{i,j} a,b,c,d,\omega_k)\, P(a,b,c,d \mid \omega_k) \quad (6)$$

and

$$g_z(k \mid \Delta_{i,j}) = P(\omega_k)\, P(x_{i,j} \mid \omega_k) \sum_{a,b,c,d} P(a,b,c,d \mid \omega_k)\, [\Pi_{m,n \in n_{i,j}} P(x_{m,n} \mid \theta_{m,n})] \quad (7)$$

and θ assumes the values 1 or 0 depending on whether the probability of X,T,L type patterns for class k is close to unity or not. Pixel (i,j) is assigned to class k for which $g(k \mid \Delta_{i,j})$ is maximum.

The configuration probabilities $P(a,b,c,d \mid \omega_k)$ are estimated from a noncontextual preclassified image. For low-resolution data, all the configurations for which $P(a,b,c,d \mid \omega_k)$ is non zero are used in evaluating the decision functions. For high-resolution data, only X,T,L type configurations are considered. The probabilities p,q,r corresponding to X,T,L type configurations are estimated from classified sample as

$$p = \frac{M_x/(M-W)}{1-W}, \qquad q = \frac{M_t/M}{1-W}, \qquad r = \frac{M_l/M}{1-W}$$

where M_x, M_t, M_l are the number of occurrences of the X,T,L type configurations, respectively, in a preclassified image: i.e.,

$$M = M_x + M_t + M_l \text{ and } W = \sum_{k=1}^{K} (P(\omega_k))^2.$$

For estimating the spatial correlations, the method proposed in Hjort *et al.* (1985) is used. It is assumed that the observation $x_{i,j}$ is composed of two independent components $y_{i,j}$ and $e_{i,j}$. $y_{i,j}$ accounts for class dependent variations while $e_{i,j}$ accounts for class independent variations. It is assumed that contextual information is in the form of spatial correlations between the $e(i,j)$ of the neighboring pixels. $x_{i,j} = y_{i,j} + e_{i,j}$, where $(y_{i,j} \mid \theta_{i,j} = \omega_k) \sim N(\mu_k, \Sigma_k)$ and $e_{i,j} \sim N(0,\Sigma_0)$. The structure of variability in $y_{i,j}$ is given by the covariance matrix Σ_k. The $e_{i,j}$ s are spatially correlated, but no correlation exists between the components of $e_{i,j}$. That is to say, the covariance matrix of $e_{i,j}$ is a diagonal matrix. If $e_{i,j}(p)$ and $e_{i,j}(q)$ denote the p-th and q-th components of $e_{i,j}$, then

$$\text{cov}\,(e_{i,j}(p), e_{m,n}(q)) = \begin{cases} 0, & \text{if } p \neq q \\ \mathbf{R}_{k,l}(q,q), & \text{if } p=q,\ k=|i-m|,\ l=|j-n| \end{cases}$$

Under these assumptions, the distribution of $(\Delta_{i,j} \mid I^*, \omega_k)$ is (for the four-neighbor case $\Delta_{i,j} = [x_{i-1,j}, x_{i,j+1}, x_{i+1,j}, x_{i,j-1}, x_{i,j}]^T$) $5p$ dimensional multivariate normal with mean vector $\boldsymbol{\mu}_\Delta = [\mu_a, \mu_b, \mu_c, \mu_d, \mu_k]^T$ and covariance matrix $\boldsymbol{\Sigma}_\Delta$ given by

$$\boldsymbol{\Sigma}_\Delta = \begin{pmatrix} \boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_0 & \mathbf{R}_{11} & \mathbf{R}_{20} & \mathbf{R}_{11} & \mathbf{R}_{10} \\ \mathbf{R}_{11} & \boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_0 & \mathbf{R}_{11} & \mathbf{R}_{02} & \mathbf{R}_{01} \\ \mathbf{R}_{20} & \mathbf{R}_{11} & \boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_0 & \mathbf{R}_{11} & \mathbf{R}_{10} \\ \mathbf{R}_{11} & \mathbf{R}_{02} & \mathbf{R}_{11} & \boldsymbol{\Sigma}_d + \boldsymbol{\Sigma}_0 & \mathbf{R}_{01} \\ \mathbf{R}_{10} & \mathbf{R}_{01} & \mathbf{R}_{10} & \mathbf{R}_{01} & \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_0 \end{pmatrix}$$

where p is the number of spectral bands considered. The parameters μ_i and Σ_i of the marginal distributions are estimated from training samples from each class k. The elements of the $\mathbf{R}$ matrices are estimated from a noncontextual preclassified image.

Methods of Accuracy Analysis

In remote sensing applications, the accuracy of an image classification refers to the extent to which classifications agree with a set of reference data. Most quantitative methods to assess classification accuracy involve an error matrix built from the two data sets: classifications and reference data. The percentage agreement uses only the main diagonal elements of the error matrix and, as such, it is a relatively simple and intuitive measure of agreement. It does not take into account the proportion of agreement between data sets that is due to chance alone, and, as such, it tends to overestimate classification accuracy (Congalton *et al.*, 1983; Rosenfield and Fitzpatric-Lins, 1986).

Congalton *et al.* (1991) reviewed the methods of accuracy assessment of land-use/land-cover classifications generated from remotely sensed data. Fienberg (1970) developed an iterative proportional fitting procedure to normalize a contingency table and include the effects of off diagonal entries on the classification accuracies. Zhuang *et al.* (1995) suggested a method for eliminating zero counts in the error matrices. It is based on the method of smoothing with pseudo counts for eliminating zero counts developed by Fienberg and Holland (1970). The approach used a Bayesian estimator to produce pseudo counts. For the examples considered here, normalization of the error matrix has been done after smoothing the error matrix with pseudo counts as suggested in Zhuang *et al.* (1995). The normalized classification results showed uniform margins and the accuracies for the individual classes.

Another important statistic frequently used in error analysis is the Kappa statistic. The Kappa (K) coefficient has come into wide use (Congalton, 1991) because it attempts to control the chance agreement by incorporating all marginal distributions of the error matrix (Cohen, 1960). Therefore, the Kappa coefficient is generally used to assess the accuracy of image classifications.

The Kappa coefficient was derived by Cohen (1960) as a flexible index for use when chance agreement between two data sets is a concern. It may be calculated as

$$K = \frac{P_0 - P_c}{1 - P_c}$$

where

$$P_0 = \sum_{i=1}^{M} P_{ii} = \frac{1}{N} \sum_{i=1}^{M} n_{ii}$$

and

$$P_c = \sum_{i=1}^{M} P_{i+}\ P_{+i} = \frac{1}{N^2} \sum_{i=1}^{M} n_{i+}\ n_{+i}$$

where N is total number of pixels, M is the number of classes, and P_{i+} and P_{+i} are the marginal distributions corresponding to the row and column. Similarly, n_{i+} and n_{+i} are the marginal totals corresponding to the rows and columns.

Testing for Significant Differences in Accuracy Coefficients

Each error matrix built for accuracy analysis must satisfy the following assumptions:

- Pixels are sampled independently,
- The different classes involved are independent mutually exclusive and exhaustive, and
- The classification runs independently.

The percentage agreement estimate parameter (P_0) follows a binomial distribution. When the sample size is large (i.e., the number of pixels used for building the error matrix, $N>100$), it follows a normal distribution. Hence, the distribution of the statistic Kappa also follows a normal distribution (Cohen, 1960). Hence, tests of significance can be performed for any given matrix to determine if the Kappa coefficient is significantly greater than zero. An approximation for the variance of Kappa (Cohen, 1960) is given by

$$\sigma_K^2 = \frac{P_0(1 - P_0)}{N(1 - P_c)^2}.$$

The significance of the difference between two Kappa coefficients can be tested using the statistic

Figure 1. Subscene 1.

Figure 2. Subscene 2.

Figure 3. Subscene 3.

$$Z = \frac{K_2 - K_1}{\sqrt{\sigma_{K_1}^2 + \sigma_{K_2}^2}} \tag{8}$$

which follows a normal distribution with zero mean and unit variance.

Test Results

The proposed method of contextual classification was tested using three examples with two data sets of the Indian Remote Sensing Satellite (IRS-1B) (subscene 1 and subscene 3) and one data set of Landsat Thematic Mapper (subscene 2). For all data sets, data from spectral bands 2, 3, and 4 were used. These data are from three different subscenes of the eastern part of India. A single band (band 4) of the images used in the examples are displayed in Figures 1, 2, and 3. Subscene 1 is 256 by 256 pixels, and subscenes 2 and 3 are 512 by 512 pixels. The dates of acquisition of subscenes 1, 2, and 3 were 22 May 1993, 18 March 1989, and 27 November 1992, respectively. No geometric corrections were made. Radiometric corrections to rectify atmospheric effects and sun angle were done by the data supplying agency (National Remote Sensing Agency, Hyderabad, India).

The details of the classes involved in these scenes and the sample sizes are presented along with the classification results. The different classes were identified by the experts in the field. Whenever distinct spectral classes were observed within a particular land-cover class, such classes were designated as class 1, class 2, etc. For example, in subscene 1, though Eucalyptus was a single class, two distinct spectral responses were observed, corresponding to the shrub types and fully grown trees. Such classes were distinguished by Eucalyptus 1 and Eucalyptus 2. The available ground truth for these classes was used for the generation of class statistics and classification accuracies. The class statistics for all the classes in terms of means and covariance matrices were estimated from the training samples to specify the noncontextual component $P(x \mid \omega_k)$ for the classes. Assuming equal *a priori* probabilities $P(\omega_k)$, the full image was classified on a per-pixel basis by the Gaussian Maximum-Likelihood (GML) classifier. In all the examples, the GML-classified image was used to estimate the contextual parameters.

The estimates of $P(I^* \mid \omega_k)$ for various configurations I^* for each class were obtained from the relative frequency of their occurrence in the GML-classified image. Any combination I^* for which $P(I^* \mid \omega_k) \leq 0.005$ was ignored and set equal to zero while using the class conditional independence

TABLE 1. CLASSIFICATION RESULTS OBTAINED WITH GML CLASSIFICATION: SUBSCENE 1

	Reference Categories														
Classification categories	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1. Sediment	13							2		14					29
2. Water tank		6	1												7
3. Deep water		2	64												66
4. Shallow water				11											11
5. Urban					54										54
6. Orchard						5									5
7. Eucalyptus-1					3		24	3		5					35
8. Eucalyptus-2								8							8
9. Degraded									10 0	13					113
10. Fallow-1										17 3					173
11. Crop-1						22					30 3	127			452
12. Crop-2												60			60
13. Fallow-2												6	27	2	35
14. Sand														159	159
Total (sample size)	13	8	65	11	57	27	24	13	10 0	20 5	30 3	193	27	161	1206

TABLE 2. CLASSIFICATION RESULTS OBTAINED WITH METHOD 1: SUBSCENE 1

Classification categories	Reference Categories 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1. Sediment	9									1					10
2. Water tank		5	4												9
3. Deep water			60												60
4. Shallow water				11											11
5. Urban					54	1	2								57
6. Orchard						24			9		17				50
7. Eucalyptus-1					2		20								22
8. Eucalyptus-2							2	7		1		1			11
9. Degraded		3				5			93	6	6	2			115
10. Fallow-1					1			12	1	19 8	2	9			223
11. Crop-1	1					2				1	27 9	14			297
12. Crop-2	1							1		4		16 6			172
13. Fallow-2													27	2	29
14. Sand	2											1		15 9	162
Total (sample size)	13	8	65	11	57	27	24	13	100	205	303	193	27	161	1206

model. The relative frequency of total occurrence of X, T, and L type configurations for each class and the relative frequency of total occurrence of other types of configurations were used for the proposed method. The GML-classified image was used in conjunction with the original image to estimate the probabilities $P_X(k),P_T(k,m),P_L(k,m)$ for each class $k = 1, 2, ..., K$ and each neighbor $m \neq k$ and the spatial correlation $\mathbf{R}_{ij}$ for spatial correlation model.

Classification of the scene was then carried out using the proposed method. If the probability of X,T,L type patterns for a particular class was found to be greater than that for other types of patterns, then the spatial correlation model was used for that class; otherwise, the class conditional independence model was used for calculating the *a posteriori* probabilities for classification. The classification accuracy was estimated using the training data by comparing the classified pixels with those in the original scene. The same training sample was used for all the classifiers in each example.

For the sake of comparison, contextual classification was also done by assuming class conditional independence (method 2) and also assuming the spatial correlation model (method 1).

An estimate of overall classification accuracy is obtained from the confusion matrix after normalization using the method of iterative proportional fitting procedure. Tables 1 to 4 present the confusion matrix for the four classification methods for subscene 1. The confusion matrices for the other scenes are not presented here for want of space. However, the normalized classification accuracy for all the methods is presented in Tables 5, 6, and 7 for subscenes 1, 2, and 3, respectively.

Example 1

The classification results of subscene 1 are shown in the form of confusion matrices in Tables 1 to 4 for the GML classification, method 1, method 2, and the proposed method. There were 14 classes. The sample sizes corresponding to each class are given by the column totals of these tables. In

TABLE 3. CLASSIFICATION RESULTS OBTAINED WITH METHOD 2: SUBSCENE 1

Classification categories	Reference Categories 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1. Sediment	9									1					10
2. Water tank		5	4												9
3. Deep water			60												60
4. Shallow water				11											11
5. Urban					54	1	2								57
6. Orchard						24			9		17				50
7. Eucalyptus-1					2		20								22
8. Eucalyptus-2							2	7		1		1			11
9. Degraded		3				5			93	6	6	2			115
10. Fallow-1					1			12	1	19 8	2	9			223
11. Crop-1	1					2				1	27 9	14			297
12. Crop-2	1							1		4		16 6			172
13. Fallow-2													27	2	29
14. Sand	2											1		15 9	162
Total (sample size)	13	8	65	11	57	27	24	13	100	205	303	193	27	161	1206

TABLE 4. CLASSIFICATION RESULTS OBTAINED WITH PROPOSED METHOD: SUBSCENE 1

Classification categories	Reference Categories 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1. Sediment	12									26					38
2. Water tank		7	2												9
3. Deep water		1	62												63
4. Shallow				11											11
5. Urban					54		6								60
6. Orchard						31			2		5				38
7. Eucalyptus-1					3		18		1						22
8. Eucalyptus-2								6		2					8
9. Degraded									98	2					100
10. Fallow-1								14	1	18	5	6			211
										5					
11. Crop-1	1					1					29	8			303
											4				
12. Crop-2	1											17			180
												9			
13. Fallow-2													27	3	30
14. Sand														15	158
														8	
Total (sample size)	13	8	65	11	57	27	24	13	10	20	30	19	27	16	1206
									0	5	3	3		1	

Table 5, we present the normalized classification accuracy for each class, the overall accuracy, the Kappa coefficient, and the variance of the Kappa coefficient (σ_λ^2). The Kappa coefficient was found to be significantly different from zero for all the methods, thereby implying that classification assignments differed significantly from random assignments. Tests of significance between the Kappa coefficients derived from the four methods were carried out as explained above using the statistic given by Equation 8. From Table 5, we note that the Kappa coefficient for method 2 is significantly high, implying thereby that method 2 is superior to the proposed method as well as to other methods.

Example 2

In this example, the subscene considered involved 15 classes. Table 6 shows the normalized classification accuracy, the kappa coefficients and the variance of Kappa coefficients for the four methods considered. The Kappa coefficients were found to be significantly different from zero. Comparing the proposed method with other methods with the help of the Z statistics given in Equation 8, we found that the proposed method and method 2 were not significantly different in terms of the Kappa coefficient. But the Kappa coefficient due to the proposed method is significantly higher than for the other two methods.

TABLE 5. NORMALIZED CLASSIFICATION ACCURACY BY DIFFERENT METHODS: SUBSCENE 1

Class	Sample size	GML	Method-1	Method-2	Proposed method
1. Sediment	13	0.9764	0.9467	0.9818	0.9259
2. Water tank	8	0.9304	0.9750	1.0000	0.9364
3. Deep water	65	0.9247	1.0000	1.0000	0.9364
4. Shallow water	11	0.9976	1.0000	1.0000	1.0000
5. Urban	57	0.9837	0.9965	1.0000	0.8788
6. Orchard	27	0.9754	0.7766	0.9936	0.9717
7. Eucalyptus-1	24	0.9741	0.8916	0.9906	0.8754
8. Eucalyptus-2	13	0.9823	0.8458	0.9793	0.8704
9. Degraded	100	0.9765	0.8388	0.9982	0.9844
10. Fallow-1	205	0.9543	0.8314	0.9841	0.7833
11. Crop-1	303	0.9262	0.9045	0.9914	0.9450
12. Crop-2	193	0.9449	0.9123	1.0000	0.9237
13. Fallow-2	27	0.9816	0.9855	1.0000	0.9934
14. Sand	161	0.9791	0.9766	1.0000	1.0000
Overall accuracy	1206	0.9648	0.9122	0.9942	0.9304
Kappa(K)		0.8031	0.8895	0.9686	0.9155
Var(K)		0.0001617	0.0000953	0.0000289	0.0000744

Example 3

In Table 7, the normalized classification accuracies, the Kappa coefficients, and the variance of the kappa coefficients are presented. Tests of significance carried out on the Kappa statistics showed significance of all Kappa values. The Kappa coefficient for the proposed method was significantly higher than that for the other methods.

Summary and Conclusions

Based on the test results, the following conclusions emerge:

- The Kappa statistics calculated from the classification results were significant, implying that all of the classifiers produced classifications significantly different from a random assignment;
- Comparisons made between the Kappa coefficients showed

TABLE 6. NORMALIZED CLASSIFICATION ACCURACY BY DIFFERENT METHODS: SUBSCENE 2

Classes	Sample size	GML	Method-1	Method-2	Proposed method
1. Forest	87	0.5595	0.8668	0.6937	0.8663
2. Hill forest	100	0.4132	0.7571	0.7080	0.7716
3. Orchard-1	178	0.9591	0.9927	0.9521	0.9943
4. Orchard-2	66	0.3857	0.8480	0.8006	0.8388
5. Sand	1107	1.0000	0.9869	0.9898	0.9880
6. Plantation	52	0.9811	0.9987	0.9129	0.9989
7. Vegetation	262	0.9163	0.8620	0.8748	0.8904
8. Fallow-1	54	0.6800	0.9708	0.9561	0.9125
9. Fallow-2	126	0.9839	0.9971	0.9622	0.9950
10. Crop-1	236	0.9954	0.9639	0.9640	0.9161
11. Crop-2	109	0.7647	0.7900	0.8246	0.7985
12. Shallow water	199	1.0000	0.9812	0.9969	0.9963
13. Deep water	48	1.0000	0.9886	0.9992	0.9959
14. Grass-1	180	0.8207	0.9014	0.8911	0.9008
15. Grass-2	195	0.7070	0.8895	0.8950	0.8857
Overall accuracy	2999	0.8111	0.9196	0.8947	0.9166
Kappa(K)		0.8692	0.8388	0.8992	0.9056
Var(K)		0.00001695	0.000056	0.00003722	0.00003506

TABLE 7. NORMALIZED CLASSIFICATION ACCURACY BY DIFFERENT METHODS: SUBSCENE 3

Classes	Sample size	GML	Method-1	Method-2	Proposed method
1. Deep water-1	1570	1.0000	1.0000	0.8668	1.0000
2. Deep water-2	584	0.9969	0.9728	0.9313	0.9987
3. Sand	445	0.6913	0.5832	0.8493	0.8090
4. Shallow water-2	457	0.8426	0.9682	0.9595	0.8946
5. Sparce vegetation	210	0.8965	0.9299	0.8308	0.9301
6. Shore vegetation	326	0.8775	0.8524	0.8621	0.9359
7. Shallow water-1	1145	0.8079	0.8861	0.8905	0.8842
8. Vegetation-1	2034	0.8694	0.8476	0.7918	0.9239
9. Shallow river water	493	0.9235	0.9039	0.9376	0.9507
10. Deep river water	309	0.9304	0.6790	0.8228	0.9242
11. Tanks & ponds	53	0.7837	0.7020	0.8759	0.8513
12. Wet land	260	0.7023	0.5747	0.8285	0.8527
13. Shore clear water	234	0.7288	0.6727	0.7444	0.8559
14. Wet ag. field	630	0.8179	0.8689	0.7944	0.8873
15. Mud flat	543	0.7452	0.7458	0.7324	0.7658
16. Dry ag. field	673	0.7552	0.8444	0.7980	0.8492
Overall accuracy	9985	0.8356	0.8145	0.8531	0.8946
Kappa(K)		0.7752	0.7293	0.7497	0.8314
Var(K)		0.00001969	0.00009716	0.000024	0.00001583

that contextual classifiers produced significantly better results than did the GML classification;

- The proposed method has been found to be superior to other methods in two out of the three examples cited above; hence, we conclude from the overall results that the proposed method has an edge over other contextual and noncontextual methods.

Acknowledgments

The authors wish to acknowledge the constructive suggestions of Dr. B. Kartikeyan, SAC Ahemmedabad, India. The help rendered by the scientists of the Regional Remote Sensing Service Center, I.I.T., Kharagpur, is gratefully acknowledged. We are also thankful to the unknown referees for their constructive criticisms. The computer code for the proposed method can be obtained from the authors at the e-mail address.

References

Amadsun, M., and R.A. King, 1998. Low level segmentation of multispectral images via agglomerative clustering of uniform neighbourhoods, *Pattern Recognition*, 21:261–268.

Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1):37–46.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment*, 37(1):35–46.

Congalton, R.G., R.G. Oderwald, and R.A. Mead, 1983. Assessing Landsat classification accuracy using discrete multivariate statistical techniques, *Photogrammetric Engineering & Remote Sensing*, 49(12):1671–1678.

DiZenzo, S., S.D. DeGloria, R. Bernstein, and H.G. Kolsky, 1987. Gaussian maximum likelihood and contextual classification algorithms for multi crop classification, *IEEE Trans. Geosci. Remote Sensing*, GE-25:805–814.

Eklundh, J.O., H. Yamamoto, and A. Rosenfeld, 1980. A relaxation method for multispectral pixel classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2:72–75.

Faugeras, O., and M. Berthod, 1981. Improving consistency and reducing ambiguity in stochastic labelling, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-3:412–423.

Fienberg, S.E., 1970. An iterative procedure for estimation in contingency tables, *The Annals of Mathematical Statistics*, 41(3):907–917.

Fienberg, S.E., and P.W. Holland, 1970. Methods for eliminating zero counts in the contingency tables, *Random Counts in Scientific Work* (G.P. Patil, editor), Pennsylvania State University Press, University Park, Pennsylvania, 1:233–260.

Haralick, R.M., and H. Joo, 1986. A contextual classifier, *IEEE Transactions on Geoscience and Remote Sensing*, GE-24:997–1007.

Haslett, J., 1985. Maximum likelihood discriminant analysis on the plane using Markovian model of spatial context, *Pattern Recognition*, 18:287–296.

Hjort, N.L., and E. Mohan, 1984. A Comparison of some contextual methods in remote sensing classification, *Proc. 18th Int. Sym. Remote Sensing Env.*, CNES, Paris.

Hjort, N.L., E. Mohn, and G. Storvick, 1985. *Contextual Classification of Remotely Sensed Data Based on an Autocorrelation Model*, Technical Report No. 13, Norwegian Computing Center.

Jhung, Y., and Philip H. Swain, 1996. Bayesian contextual classification based on modified M-estimates and Markov random fields, *IEEE Transactions on Geoscience and Remote Sensing*, GE-34: 67–75.

Kartikeyan, B., B. Gopalakrishna, M.H. Kalubarme, and K.L. Majumdar, 1994. Contextual techniques for classification of high and low resolution remote sensing data, *Int. J. Remote Sensing*, 15: 1037–1051.

Khazenie, N., and M.M. Crawford, 1990. Spatial-temporal autocorrelated model for contextual classification, *IEEE Transactions on Geoscience and Remote Sensing*, GE-28:529–539.

Landgrebe, D.A., 1980. The development of a spectral-spatial classifier for earth observational data, *Pattern Recognition*, 12:165–175.

Merical, M.B., J.C. Landgrebe, and S.S. Shen, 1984. A Spatial processing algorithm to reduce the effects of mixed pixels and increase the separability between classes, *Pattern Recognition*, 17: 525–533.

Peleg, S., 1980. A new probabilistic relaxation scheme, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2: 362–369.

Richards, J.A., D.A. Landgrebe, and P.H. Swain, 1982. A means for utilizing ancillary information in multispectral classification, *Remote Sensing of Environment*, (12):463–477.

Rosenfield, G.H., and K. Fitzpatrick-Lins, 1986. A coefficient of agreement as a measure of thematic classification accuracy, *Photogrammetric Engineering & Remote Sensing*, 52(2):223–227.

Swain, P.H., and S.M. Davis, 1978. Remote Sensing: *The Quantitative Approach*, McGraw Hill, Series.

Swain, P.H., S.B. Vardeman, and J.C. Tilton, 1981. Contextual classification of multispectral image data, *Pattern Recognition*, 13: 429–441.

Welch, J.R., and K.G. Salter, 1971. A context algorithm for pattern recognition & image interpretation, *IEEE Trans. Sys. Man. Cybern.*, SMC-1:24–30.

Xin Zhuang, Bermard Engel, Xlaoping Xlong, and Chris J. Johannsen, 1995. Analysis of classification results of remotely sensed data and evaluation of classification algorithms, *Photogrammetric Engineering & Remote Sensing*, 61(4):427–433.

(Received 28 May 1996; revised and accepted 10 April 1997; revised 22 July 1997)