CONTEXTUAL CLASSIFICATION OF MULTISPECTRAL IMAGE DATA: AN UNBIASED ESTIMATOR FOR THE CONTEXT DISTRIBUTION

JAMES C. TILTON AND PHILIP H. SWAIN
Purdue University

STEPHEN B. VARDEMAN
Iowa State University

ABSTRACT

Recent investigations have demonstrated the effectiveness of a contextual classifier that combines spatial and spectral information employing a general statistical approach. This statistical classification algorithm exploits the tendency of certain ground-cover classes to occur more frequently in some spatial contexts than in others. Indeed, a key input to this algorithm is a statistical characterization of the context: the context distribution. Here we discuss an unbiased estimator of the context distribution which, besides having the advantage of statistical unbiasedness, has the additional advantage over other estimation techniques of being amenable to an adaptive implementation in which the context distribution estimate varies according to local contextual information. Results from applying the unbiased estimator to the contextual classification of three real LANDSAT data sets are presented and contrasted with results from non-contextual classifications and from contextual classifications utilizing other context distribution estimation techniques.

I. INTRODUCTION

The machine classification of multispectral image data collected by remote sensing devices aboard aircraft and spacecraft has usually been performed such that each pixel (picture element) is classified individually and independently. The information used by this classifier is only spectral or, in some cases, spectral and temporal. There is no provision for using the spatial information inherent in the data. In contrast, when scanner data are displayed in image form, a human analyst routinely uses spatial information to establish a context for deciding what a particular pixel in the imagery might be. Using this context together with spectral information, the analyst may easily identify roads, delineate boundaries of agricultural fields, and differentiate between grass in an urban setting (e.g., lawns) and grass in an agricultural setting (e.g., pasture or forage crops) where a point-by-point classifier utilizing spectral information alone would have much difficulty in doing so.

The ECHO (Extraction and Classification of Homogeneous Objects) process is a variety of contextual classifier which has been found useful for classifying data sets which contain homogeneous objects that are large compared to the resolution of the imagery. This classifier cannot be used effectively, however, if the data set does not contain a significant number of these large homogeneous objects.

In several recent papers, we have described a general statistical classification method for exploiting both spatial and spectral information when classifying multispectral image data. This contextual classifier exploits the tendency alluded to earlier of certain ground-cover classes to occur more frequently in some contexts than in others. Unlike the ECHO process, this classifier can be used to advantage on any data set, even those data sets that do not have identifiable homogenous objects, such as is generally the case in forested, urban and other inhomogeneous areas.

We shall briefly review the statistical basis of the contextual decision rule and earlier methods for estimating a statistical characterization of context: the context distribution. We will then describe an unbiased estimator of the context distribution. Besides having the advantage of statistical unbiasedness, this estimator has the additional advantage over other estimation techniques of being amenable to an adaptive implementation in which the context distribution estimate varies according to local contextual information. Results from applying the unbiased estimator to the contextual classification of three real LANDSAT data sets are then presented and contrasted with results from non-contextual classifications and from contextual classifications utilizing other context distribution estimation techniques.

II. THEORETICAL BASIS OF THE CLASSIFIER

Consistent with the general characteristics of imaging systems for remote sensing, we assume a two-dimensional array of \( N = N_1 \times N_2 \) random observations.
(pixels) $X_{ij}$ having fixed but unknown classification $\vartheta_{ij}$, as shown in Figure 1. The observation $X_{ij}$ consists of $n$ measurements (usually containing spectral and/or temporal information), while the classification $\vartheta_{ij}$ can be any one of $m$ spectral or information classes* from the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$.

$$\vartheta_{11}, \vartheta_{12}, \ldots, \vartheta_{1N_2}$$
$$\vartheta_{21}, \vartheta_{22}, \ldots, \vartheta_{2N_2}$$
$$\vdots$$
$$\vartheta_{N_1,1}, \ldots, \vartheta_{N_1,N_2}$$

Figure 1. A two-dimensional array of $N = N_1 \times N_2$ pixels.

Let $X$ denote a vector whose components are the random observations:

$$X = \{X_{ij} | i = 1,2, \ldots, N_1; j = 1,2, \ldots, N_2\}^T.$$  

Similarly, let $\vartheta$ be the vector of states (true classifications) associated with the observations:

$$\vartheta = \{\vartheta_{ij} | i = 1,2, \ldots, N_1; j = 1,2, \ldots, N_2\}^T.$$  

The following notation will be useful. Let $\vartheta^p \in \Omega^P$ and $X^p \in (R^n)^P$ stand for $p$-vectors of classes and $n$-dimensional measurements, respectively; each component of $\vartheta^p$ is a variable which can take on any classification value; each component of $X^p$ is a $n$-dimensional random vector which can take on values in the observation space.

Let the action (classification) taken with respect to pixel $(i,j)$ be denoted by $a_{ij} \in \Omega$. We restrict the action $a_{ij}$ to be a function of a specified subset of observations in $X$. This subset includes, along with $X_{ij}$, $p-1$ observations spatially near to, but not necessarily adjacent to, $X_{ij}$. These $p-1$ observations serve as the spatial context for $X_{ij}$ and are taken from the same spatial positions relative to pixel position $(i,j)$ for all $i$ and $j$. Call this arrangement of pixels together with $X_{ij}$ the $p$-context array, several examples of which are shown in Figure 2. Group the $p$ observations in the $p$-context array into a vector of observations $X^P = (X_{i1}, X_{i2}, \ldots, X_{ij})^T$ and let $\vartheta^p$ be the vector of true but unknown classifications associated with the observations in $X^p$. Note that the $\vartheta^p$ and $X^p$ are the particular instance of $\vartheta^p$ and $X^p$ associated with pixel position $(i,j)$. Correspondence of the components of $\vartheta^p$, $X^p$, $\vartheta$ and $X$ to the positions in the $p$-context array is fixed but arbitrary except that the $p^a$ components will always correspond to the pixel to be classified.

Let the loss suffered by taking action $a_{ij}$ be denoted by $\lambda(\vartheta_{ij}, a_{ij})$ for some fixed non-negative function $\lambda(\cdot, \cdot)$. The expected average loss (or risk) suffered

$$\begin{array}{c|c|c}
\text{i-1,j+1} & \text{i,j} & \text{i,j+2} \\
\hline
\text{i,j-1} & \text{i,j} & \text{i,j+1} \\
\text{i+1,j} & \text{i,j-1} & \text{i,j} \\
\end{array}$$

a $p=2$ choice

a $p=2$ choice

Figure 2. Examples of $p$-context arrays.

over the $N$ classifications in the classification array is

$$R_{\vartheta} = E \left[ \frac{1}{N} \sum_{i \neq j} \lambda(\vartheta_{ij}, a_{ij}(X_{ij})) \right]$$  \hspace{1cm} (1)

where the expectation is with respect to the distribution of $X$.

Now consider finding a decision rule of the form

$$a_{ij}(X_{ij}) = d(X_{ij})$$  \hspace{1cm} (2)

for a fixed function $d(\cdot)$ mapping $p$-vectors of observations to actions so that $R_{\vartheta}$ is minimized. If we require that the distributions of the $X_{ij}$ are spatially invariant, i.e. the value of the probability density for $X_{ij}$ depends only on the measurement values in $X_{ij}$ and the set of classifications in $\vartheta_{ij}$ and not the location $(i,j)$, the risk, $R_{\vartheta}$, can be written as

$$R_{\vartheta} = \sum_{\vartheta^p \in \Omega^P} G(\vartheta^p) \int \lambda(\vartheta, a) f(X_{ij} | \vartheta^p) dX^P$$

$$= \int \sum_{\vartheta^p \in \Omega^P} G(\vartheta^p) \lambda(\vartheta, a) f(X_{ij} | \vartheta^p) dX^P$$  \hspace{1cm} (3)

where $G(\vartheta^p)$, the context distribution, is the relative frequency with which $\vartheta^p$ occurs in the array $\vartheta$, and $a_{\vartheta^p}$ is the $p^a$ element of $\vartheta^p$. For any array $\vartheta$, a decision rule $d(X_{ij})$ minimizing $R_{\vartheta}$ can be obtained by minimizing the integrand in (3) for each $X_{ij}$; thus for a specific $X_{ij}$ (an instance of $X^P$), an optimal action is:

$$d(X_{ij}) = \text{the action (classification) $a$ which minimizes}$$

$$\sum_{\vartheta^p \in \Omega^P} G(\vartheta^p) \lambda(\vartheta^p, a) f(X_{ij} | \vartheta^p).$$  \hspace{1cm} (4)

In practice, a "0-1 loss function" is usually assumed, i.e.,

$$\lambda(\vartheta, a) = \begin{cases} 0, & \text{if } \vartheta = a \\ 1, & \text{if } \vartheta \neq a \end{cases}.$$

* Spectral classes are spectrally differentiable subclasses of information classes (the classes of interest).
Then (4) simplifies and the decision rule becomes:

\[ d(X_d) = \text{the action } a \text{ which maximizes } \sum_{\varphi \in \mathcal{CP}} C(\varphi) f(X_d | \varphi_a). \]  

(5)

We now assume class-conditional independence for the observations. This assumption means that the joint class-conditional density over the p-context array can be written as

\[ f(X_d | \varphi) = \prod_{k=1}^{p} f(X_k | \varphi_k) \]  

(6)

where \( X_k \) and \( \varphi_k \) are the \( k^{th} \) elements of \( X_d \) and \( \varphi \), respectively. Evidence that this is a reasonable assumption may be found in Yamamoto. With this assumption, the decision rule in (5) becomes:

\[ d(X_d) = \text{the action } a \text{ which maximizes } \sum_{\varphi \in \mathcal{CP}} C(\varphi) \prod_{k=1}^{p} f(X_k | \varphi_k). \]  

(7)

A more detailed derivation of this decision rule can be found in Swain, et al.1

The optimal choice of \( d(.) \) cannot be implemented in practice since it depends on \( C(\varphi) \) and the \( f(X_k | \varphi_k) \) which are unknown. Methods for estimating the \( f(X_k | \varphi_k) \) are well established from considerable experience in using the conventional non-contextual maximum likelihood decision rule.3 When the classification set \( \mathcal{G} \) consists of spectral classes, the \( f(X_k | \varphi_k) \) are assumed to be multivariate normal densities. In the case where the classification set \( \mathcal{G} \) consists of information classes, the \( f(X_k | \varphi_k) \) are assumd to be weighted sums of multivariate normal densities. We will next discuss methods for estimating the context distribution, \( C(\varphi) \).

III. CONTEXT DISTRIBUTION ESTIMATION: EARLIER TECHNIQUES

Simulated data sets were utilized in the earliest experiments exploring the effectiveness of classifying multispectral remote sensing data using context classification as defined by the set of discriminant functions in (7). This was done to demonstrate the effectiveness of the classifier given that the underlying assumptions in the classification model are satisfied. At first, the context distribution was found by simple tabulation from the true classification used as a template for the data simulation. As reported in Swain, et al.1 the classifier was very effective when the context distribution was determined in this way.

When dealing with real data, there is no direct way of determining the context distribution. We cannot tabulate the context distribution from the true classification since the true classification is not known. However, we do expect that, at least for large

\[ N = N_1 \times N_2, \]  

the decision rule in (5) where \( C(\varphi) \) is replaced by an estimate \( C(\varphi) \) based on the data, \( X \), will have risk \( R_g \), approximating that of the optimal rule. Thus we should be able to base an adequate estimate of the context distribution on the data or, more practically, on representative sections from the data designated as a training set. The most straightforward way to develop an estimate of the context distribution from the training set would be to perform a conventional non-contextual classification of the training set and use the context distribution as tabulated from this classification as an estimate of the context distribution. One could then further refine this estimate of the context distribution by making another estimate from the contextual classification, and even iterate in this way until no further improvement in classification accuracy was obtained.

This iterative "classify-and-count" method was tested on one simulated data set and two real data sets. As reported in Swain, et al.1 this method gave excellent results on the simulated data set, but disappointing results on the real data sets, stimulating a search for alternative methods for estimating the context distribution. One such method is the ground-truth-guided method. In this method, roughly equal subsets of the ground truth data are designated as a training set for estimating the context distribution and a test set for evaluating the classification results. The ground truth data are, of course, represented in terms of information classes. When the estimation is to be done in terms of spectral classes rather than information classes, the following method is used:

1. Perform a conventional non-contextual classification of the training set using uniform prior probabilities, but allow the the classifier to choose only among spectral classes associated with the information class designated by the ground truth.
2. Estimate the context distribution by tabulation from the resulting 100-percent accurate classification of the training set.
3. Classify the entire scene with the contextual classifier and evaluate the results over a test set disjoint from the training set.

When the estimation is to be done in terms of information classes, the restricted spectral class classification in step 1 above must still be performed. In this case, however, this classification is used to provide (by tabulation) an estimate of the weights used in the weighted sum of class-conditional normal densities that make up the set of densities \( f(X_k | \varphi_k) \) in (7). Each weight is the relative frequency of occurrence in the training set of a particular spectral class for a given information class. The entire scene is then classified in terms of information classes using the contextual classifier, and evaluated over a test set disjoint from the training set, as in the spectral class case.

Both the spectral and information class formulations of the ground-truth-guided method were tested on two 50-pixel-square Landsat data sets. One data set was a LACIE data set from Hodgeman County, Kansas, containing pasture, wheat, corn and fallow fields. The other data set was from Tippecanoe County, Indiana, containing residential and commercial areas in northern Lafayette and West Lafayette, Indiana, as well as...
areas of forest, agriculture and water (the Wabash River). For both data sets, the restricted spectral class classification was performed over the first 25 lines of the data set and the context distribution was estimated over those 25 lines. Contextual classifications of the scenes were performed and classification accuracies* were evaluated over the last 25 lines as well as over the entire data set.

Tables 1 and 2 present the results from contextual classifications using four-nearest-neighbor (4nn) estimates of the context distribution (the p=5 choice in Figure 2) for both the spectral and information class formulations of the ground-truth-guided method (gtgm). These results are also compared to the accuracies obtained from uniform-priors and estimated-priors non-contextual maximum likelihood classifications. The prior probabilities for the estimated-priors non-contextual classifications were estimated by tabulation from the uniform-priors non-contextual classification. These results show that contextual classifications using the ground-truth-guided method for estimating the context distribution give significantly better results than non-contextual classifications on these data sets. For these cases, the spectral class formulation of the ground-truth-guided method generally produces higher classification accuracies. However, since the spectral class estimate of the context distribution has substantially more non-zero elements than the information class estimate, contextual classifications using the spectral class formulation generally take over twice the computer time required for the information class formulation.

While this method can produce good estimates of the context distribution, it suffers the limitation that it requires large areas of spatially contiguous ground truth data. When such detailed ground truth data are not available, some other method is needed.

*Classification accuracy can be tabulated in two ways. Overall accuracy is simply the overall number of correct classifications divided by the total number attempted. Average-by-class accuracy is obtained by first computing the accuracy for each class and then taking the arithmetic average of the class accuracies. The latter is significant when the classification results exhibit a tendency to discriminate in favor of or against a subset of the classes.

Table 1. Comparison of the contextual classifier using the ground-truth-guided method with non-contextual classifiers; Hodgeman County, Kansas, Landsat Data Set.

<table>
<thead>
<tr>
<th>Classification</th>
<th>% Accuracy lines 26-50</th>
<th>% Accuracy lines 1-50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>Average-by-Class</td>
</tr>
<tr>
<td>uniform priors</td>
<td>81.5</td>
<td>78.2</td>
</tr>
<tr>
<td>estimated priors</td>
<td>82.2</td>
<td>78.3</td>
</tr>
<tr>
<td>4nn gtgm, spectral</td>
<td>85.4</td>
<td>81.6</td>
</tr>
<tr>
<td>4nn gtgm, information</td>
<td>85.3</td>
<td>81.4</td>
</tr>
</tbody>
</table>

Table 2. Comparison of the contextual classifier using the ground-truth-guided method with non-contextual classifiers; Tippecanoe County, Indiana, Landsat Data Set.

<table>
<thead>
<tr>
<th>Classification</th>
<th>% Accuracy lines 26-50</th>
<th>% Accuracy lines 1-50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>Average-by-Class</td>
</tr>
<tr>
<td>uniform priors</td>
<td>82.7</td>
<td>81.7</td>
</tr>
<tr>
<td>estimated priors</td>
<td>84.2</td>
<td>82.0</td>
</tr>
<tr>
<td>4nn gtgm, spectral</td>
<td>88.7</td>
<td>91.1</td>
</tr>
<tr>
<td>4nn gtgm, information</td>
<td>88.2</td>
<td>87.3</td>
</tr>
</tbody>
</table>
The "Power Method" was the next method investigated as a generally applicable method of estimating the context distribution. To employ the method, one raises the relative frequency count for each class configuration to a power and uses the result as the context distribution estimate. This method is described in detail in Tilton, et al. The context distribution estimates generated by the Power Method can produce classification accuracies of roughly the same high level as produced by the ground-truth-guided method. However, the method is very inconvenient to use.

With the the Power Method, an estimate of the context distribution is tabulated from a uniform-priors non-contextual classification of the training set. Then contextual classifications of the training set and test set are performed using a power of the tabulated context distribution. To achieve the best possible results, a second iteration of this procedure must generally be performed, using a context distribution estimate tabulated from the training set of the first iteration classification. Unfortunately, no reliable predictor has been found for the optimal power to be used for the first or second iteration. It is not even the case that the most accurate first iteration classification will provide in general the best template for the second iteration. Further, on certain data sets, a spectral-class context-distribution estimate produces the best results, while on other data sets an information-class formulation works better. Despite the good results possible with the Power Method, these ambiguities make this method difficult to use, and not useful for practical applications. A search for a better generally applicable method for estimating the context distribution has led to the unbiased estimation technique described in the next section.

IV. CONTEXT DISTRIBUTION ESTIMATION: UNBIASED ESTIMATOR

One tactic for seeking an optimal estimate of the context distribution, \( C(\mathbf{x}^p) \), is to look for an estimator function, \( T_{\mathbf{x}^p}(\mathbf{x}) \), which minimizes the mean-squared error given by

\[
MSE = E\left[ T_{\mathbf{x}^p}(\mathbf{x}) - C(\mathbf{x}^p) \right]^2. \tag{8}
\]

Equation (8) can be rewritten as

\[
MSE = \text{Var}[T_{\mathbf{x}^p}(\mathbf{x})] + b^2 \tag{9}
\]

where \( \text{Var}[T_{\mathbf{x}^p}(\mathbf{x})] \) is the variance of the estimate \( T_{\mathbf{x}^p}(\mathbf{x}) \) and \( b \) is the bias given by

\[
b = E[T_{\mathbf{x}^p}(\mathbf{x})] - C(\mathbf{x}^p). \tag{10}
\]

Finding the minimum mean-squared-error estimate is generally a difficult task, but since bias represents a systematic error, a reasonable approach would be to control bias before considering the variance. The best one can do in controlling bias is to seek an unbiased estimator, i.e., one for which \( b = 0 \).

As we saw in the previous section, the classify-and-count method performed poorly in tests on real Landsat data sets. One reason for this is that the estimate can be statistically biased. To prove this, consider the classification model as presented in section II. In addition to the symbol definitions given there, we make the following definitions. Let \( \mathbf{\hat{y}} \) be the vector of classifications

\[
\mathbf{\hat{y}} = [\hat{y}_i | i = 1, 2, \ldots, N; j = 1, 2, \ldots, N_j]^T
\]

where \( \hat{y}_i \) is the classification estimate from a non-contextual classification of the observation \( X_{ij} \). Let \( \mathbf{\hat{y}} \) be a p-vector of classification estimates associated with the observations in the p-context array, \( X_{ij} \). Similarly, let \( \mathbf{\hat{y}}^p \) be such an estimate associated with an arbitrary p-context array, \( X_{ij}^p \). Let \( \mathbf{p} \in \mathbf{p}^p \) represent an arbitrary p-vector of classes. The classify-and-count method can be described by the following estimator function for \( C(\mathbf{x}^p) \):

\[
T_{\mathbf{x}^p}(\mathbf{x}) = \mathbf{\hat{y}}(\mathbf{x}^p) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N_j} I(X_{ij}, \mathbf{p}) \tag{11}
\]

where

\[
I(X_{ij}, \mathbf{p}) = \begin{cases} 1, & \text{if } \mathbf{p} = \mathbf{\hat{y}}_i \\ 0, & \text{otherwise.} \end{cases}
\]

The expected value of \( T_{\mathbf{x}^p}(\mathbf{x}) \) is then

\[
E[T_{\mathbf{x}^p}(\mathbf{x})] = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N_j} E[I(X_{ij}, \mathbf{p})] = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N_j} \sum_{\mathbf{p}^p \subset \mathbf{p}^p} E[I(X_{ij}, \mathbf{p}^p)]
\]

\[
= \frac{1}{N} \sum_{\mathbf{p}^p \subset \mathbf{p}^p} \sum_{\mathbf{p}^p} \sum_{\mathbf{p}^p} E[I(X_{ij}, \mathbf{p}^p)]
\]

\[
= \sum_{\mathbf{p}^p \subset \mathbf{p}^p} C(\mathbf{p}^p) \int_{\mathbf{p}^p \subset \mathbf{p}^p} f(\mathbf{p}^p | \mathbf{x}^p) d\mathbf{x}^p. \tag{12}
\]

Equations (10) and (12) show that the bias of the classify-and-count method is the difference between \( C(\mathbf{x}^p) \) and a weighted sum of \( C(\mathbf{x}^p) \). Note that this bias is independent of \( N \), and cannot be reduced by increasing sample size. The bias can be non-zero or zero, depending on the values of \( C(\mathbf{x}^p) \) and integrals in (12). To show this explicitly, let's consider the simple special case of a two-class problem (\( m = 2 \)) estimating non-contextual relative frequencies of classes (\( p = 1 \)) for univariate random observations (\( n = 1 \)). Let the non-contextual classifier used to produce \( \mathbf{\hat{y}} \) be the uniform-priors maximum-likelihood classifier with the decision rule:

\[
d(X_{ij}) = \text{the action } a \text{ which maximizes } f(X_{ij} | a)
\]

for all \( a \in \{a_1, a_2\} \). The densities, \( f(X_{ij} | a) \), are assumed to be normal with mean and variance \( \mu_a = -1 \) and \( \sigma_a^2 = 1 \) for class \( a_1 \) and mean and variance \( \mu_2 = 1 \) and \( \sigma_2^2 = 1 \) for class \( a_2 \). For class \( a_1 \), we have:

\[
E[T_{a_1}(X)] = \sum_{k=1}^{2} C(a_k) \int_{a_k} f(X | a_k) dX
\]

\[
= \sum_{k=1}^{2} C(a_k) \int_{a_k} f(X | a_k) dX
\]

1981 Machine Processing of Remotely Sensed Data Symposium
\[
E[T_{xy}(X)] = 0.84G(\omega_1) + 0.84G(\omega_2).
\]

We have shown, then, that the classify-and-count method does indeed produce unbiased estimates of the context distribution.

The unbiased estimator we have adopted can be most easily described by first considering the p=1 case and then generalizing to the arbitrary p-context array. For p=1, we examine the equation

\[
\int h_k(X) \left[ \sum_{i=1}^{m} f(X | \omega_i) G(\omega_i) \right] dX = \sum_{i=1}^{m} \left[ \int h_k(X) f(X | \omega_i) dX \right] G(\omega_i)
\]

where \( m \) is the number of classes; \( f(X | \omega_i) \), \( i=1,2,\ldots,m \), are the class-conditional densities described earlier; and the functions \( h_k(X) \), \( k=1,2,\ldots,m \), can be any set of \( m \) linearly independent functions. Equation (15) is valid provided all indicated sums and integrals are well defined, which will, for example, be the case when all of the functions in (15) are bounded. The functions \( G(\omega_i) \) and \( f(X | \omega_i) \) are always bounded because \( G(\omega_i) \) is a relative frequency function and \( f(X | \omega_i) \) is a multivariate normal density function. The functions \( h_k(X) \) considered in the following development will also always be bounded.

The left-hand side of (15), which looks like the expected value of \( h_k(X) \), can be estimated from the data \( X \) as follows:

\[
\int h_k(X) \left[ \sum_{i=1}^{m} f(X | \omega_i) G(\omega_i) \right] dX = \frac{1}{N} \sum_{i=1}^{m} \left[ \int h_k(X) f(X | \omega_i) dX \right] G(\omega_i)
\]

where \( N \), \( N_1 \) and \( N_2 \) are as defined in Figure 1, and \( k \in \{1,2,\ldots,m\} \). Applying (15) and (16) \( m \) times, one for each class, we can write

\[
\begin{bmatrix}
  h_1(X) \\
  h_2(X) \\
  \vdots \\
  h_m(X)
\end{bmatrix} = \begin{bmatrix}
  I_{11} & I_{12} & \cdots & I_{1m} \\
  I_{21} & I_{22} & \cdots & I_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  I_{m1} & I_{m2} & \cdots & I_{mm}
\end{bmatrix} \begin{bmatrix}
  G(\omega_1) \\
  G(\omega_2) \\
  \vdots \\
  G(\omega_m)
\end{bmatrix}
\]

where

\[
I_{kl} = \int h_k(X) f(X | \omega_l) dX.
\]

This can be more succinctly represented in vector-matrix notation as

\[
\hat{C} = I C.
\]

Now \( C \) can be estimated by solving

\[
\hat{C} = I^{-1} h \triangleq T
\]

where \( T = (T_1(X),T_2(X),\ldots,T_m(X))^T \) is the vector equivalent of \( I(X) \) in (5), (8) and (10).

To show that \( T \) is indeed an unbiased estimator for \( C \), we note that

\[
E(T) = E(I^{-1} h) = I^{-1} E(h).
\]

Looking at \( E(h) \) element by element we have

\[
E \left[ \frac{1}{N} \sum_{i=1}^{N_1} h_k(X_{\omega_i}) \right] = \frac{1}{N} \sum_{i=1}^{N_1} E \left[ h_k(X_{\omega_i}) \right]
\]

Thus

\[
E(h) = I C
\]

and (20) becomes

\[
E(T) = I^{-1} E(h) = I^{-1} I C = C
\]

proving that \( T \) is an unbiased estimator for \( C \).

It is convenient to use a function of the class-conditional densities for the functions \( h_k(X) \). More specifically, let \( h_k(X) = (2\pi)^{\frac{n}{2}} f(X | \omega_k) \) and write (17b) as

\[
I_{kl} = (2\pi)^{\frac{n}{2}} \int f(X | \omega_k) f(X | \omega_l) dX
\]

where \( n \) is the dimensionality of \( X \). Assuming the \( \omega_k \) are normally distributed spectral classes with respective mean vectors \( \mu_k \) and covariance matrices \( \Sigma_k \) \( (k=1,2,\ldots,m) \), we find

\[
I_{kl} = \frac{1}{2} \left[ \text{det}(\Sigma_k + \Sigma_l) \right]^{\frac{1}{2}} \exp \left[ -\frac{1}{2} (\mu_k - \mu_l)^T (\Sigma_k + \Sigma_l)^{-1} (\mu_k - \mu_l) \right].
\]

When the \( \omega_k \) are information classes, the \( I_{kl} \) are weighted sums of terms of the form given in (23).
When the estimate is made in terms of information classes, estimates must be made of the weights used to form the weighted sum of the class-conditional normal densities of the spectral subclasses. For each information class, the weights are estimated by using the unbiased estimator with p=1 for the spectral classes which make up the information class being considered.

The calculation of the estimate of $G$ can proceed in one of two alternative ways. The vector $A$ can be calculated for the entire image (as in (17a)), then multiplied by $I^{-1}$ to give $T \equiv G$; or as the $A_k(x_i)$ are calculated at each data point (pixel), the product with $I^{-1}$ can be performed. The average of this product over the entire image is then $T \equiv G$. The methods are completely equivalent; the difference between them amounts to a change in order of summation. However, the second method must be used when this unbiased estimator is extended to the arbitrary p-context array case, because the use of the first method for large values of p would require an impractical amount of storage. In calculating the estimate of $C(\phi, \psi)$ at each image data point using the second method, individual unbiased estimates of the prior probabilities of each class are made for each position in the p-context array, and cross-products of these prior probabilities are taken to form the unbiased estimate of $C(\phi, \psi)$ based on that image point. To save computer storage space, the cross-products having values below a specified threshold are ignored. The estimate of $C(\phi, \psi)$ for the entire image is the average of the estimates of $C(\phi, \psi)$ based on all the individual image points in the scene.

The unbiased estimator can be modified to provide an adaptive estimator of the context distribution. The local context distribution estimate for a particular $n_1 \times n_2$ block of image data is made from a $m_1 \times n_2$ block ($m_1 \geq n_1$ and $m_2 \geq n_2$). The $n_1 \times n_2$ block of image data is then classified using this local estimate of the context distribution. This process is repeated until the entire data set is classified. Better results have generally been obtained when $m_1 \geq n_1$ and $m_2 \geq n_2$. If $m_1 = n_2$ and $m_2 = n_2$, the context distribution estimate is not accurate for the pixels at the edges of the image data block being classified. Tests on three 50-pixel-square Landsat data sets have indicated good choices for $n_1$ and $n_2$ ranging from 10 up to 25 with the corresponding choices for $m_1$ and $m_2$ being 8 to 10 larger than the values chosen for $n_1$ and $n_2$.

VI. CONCLUDING REMARKS

It had been shown earlier in this research\(^1\) that the contextual classifier can provide improved classification performance, as compared to non-contextual classification, when accurate characterization of the context distribution are available. The ground-truth-guided method has been shown to provide sufficiently accurate estimates of the context distribution, but suffers the disadvantage of requiring sizeable amounts of spatially contiguous ground truth. The unbiased estimator described herein overcomes this disadvantage, providing good estimates of the context distribution while requiring no more ground truth data than is required for a non-contextual classification. Furthermore, the unbiased estimator is amenable to an adaptive implementation so that the resulting context distribution estimate is more closely tailored to local conditions in the image data.

REFERENCES


---

**Table 3. Comparison of the contextual classifier using various unbiased estimator formulations and the uniform-priors non-contextual classifier.**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Classification</th>
<th>%Accuracy</th>
<th>Average- by-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td></td>
</tr>
<tr>
<td>Hodgeman County,</td>
<td>uniform-priors non-contextual</td>
<td>82.0</td>
<td>75.9</td>
</tr>
<tr>
<td>Kansas, 50-pixel-square Landsat</td>
<td>4nn unbiased, spectral class whole image est.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(evaluated over lines and columns 6 through 50)</td>
<td>(nonadaptive)</td>
<td>83.1</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, spectral class adaptive est.,</td>
<td>84.0</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td>25x25 from 25x25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, information class adaptive est.,</td>
<td>84.0</td>
<td>75.0</td>
</tr>
<tr>
<td></td>
<td>25x25 from 35x35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monroe County, Indiana, 50-pixel-square Landsat</td>
<td>uniform-priors non-contextual</td>
<td>83.1</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, spectral class whole image est.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(nonadaptive)</td>
<td>84.4</td>
<td>84.4</td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, spectral class adaptive est.,</td>
<td>84.3</td>
<td>83.9</td>
</tr>
<tr>
<td></td>
<td>25x25 from 25x25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, information class adaptive est.,</td>
<td>88.9</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td>17x17 from 25x25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tippecanoe County,</td>
<td>uniform-priors non-contextual</td>
<td>81.8</td>
<td>83.4</td>
</tr>
<tr>
<td>Indiana, 50-pixel-square Landsat</td>
<td>4nn unbiased, spectral class whole image est.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(nonadaptive)</td>
<td>86.2</td>
<td>87.9</td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, spectral class adaptive est.,</td>
<td>86.7</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>25x25 from 25x25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, information class adaptive est.,</td>
<td>86.2</td>
<td>89.1</td>
</tr>
<tr>
<td></td>
<td>25x25 from 25x25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4nn unbiased, information class adaptive est.,</td>
<td>86.9</td>
<td>89.7</td>
</tr>
<tr>
<td></td>
<td>10x10 from 20x20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*1981 Machine Processing of Remotely Sensed Data Symposium*
Figure 3. Visual comparison of classification results, Tippecanoe County, Indiana, Landsat data set. (a) uniform-priors no-context, (b) estimated-priors no-context, and (c) four-nearest-neighbor adaptive (17x17 from 27x27) unbiased estimator (d) reference image.