Statistical Modeling for Improved Land Cover Classification

Yunxin Zhao, Xiaobo Zhou, K. Palaniappan, and X. Zhuang

Dept. of Computer Engineering and Computer Science, Univ. of Missouri-Columbia

ABSTRACT

Novel statistical modeling and training techniques are proposed for improving classification accuracy of land cover data acquired by LandSat Thermatic Mapper (TM). The proposed modeling techniques consist of joint modeling of spectral feature distributions among neighboring pixels and partial modeling of spectral correlations across TM sensor bands with a set of semi-tied covariance matrices in Gaussian mixture densities (GMD). The GMD parameters and semi-tied transformation matrices are first estimated by an iterative maximum likelihood estimation algorithm of Expectation-Maximization, and the parameters are next tuned by a minimum classification error training algorithm to enhance the discriminative power of the statistical classifiers. Compared with a previously proposed single-pixel based Gaussian mixture density classifier, the proposed techniques significantly improved the overall classification accuracy on eight land cover classes from imagery data of Missouri state.

Keywords: Land cover, LandSat TM data, Gaussian mixture density, semi-tied covariance, minimum classification error training, fusion

1. INTRODUCTION

Classification of land cover classes based on multispectral data acquired by Landsat Thermatic Mapper (TM) has been an active area of research in remote sensing. Several noncontextual classifiers have been proposed previously that categorize imagery pixels based on spectral measurements at the individual pixels while ignoring contextual information in neighboring pixels. Examples include statistical classifiers of Gaussian [1] and Gaussian mixture [2] models, and piecewise linear classifier of binary decision trees [3]. In [1], a Bayesian contextual classifier was proposed to overcome the deficiency of noncontextual classifiers, and robust statistic method of M-estimate was proposed for parameter estimation of pixel-based Gaussian spectral distributions. In [4], a linear canonical discriminative analysis was proposed to maximize land cover class discrimination given a fixed dimension of textual feature representation [4]. In the current work, novel statistical modeling and training techniques are proposed for improving classification accuracy of TM land cover data over a baseline noncontextual classifier of Gaussian mixture models [2]. The classification task consists of eight land cover classes, the major classes being herbaceous, forest, and woodland.

In the method of [2], pixel-based Gaussian mixture density (GMD) modeling with diagonal covariance matrices was used. This approach was ineffective in capturing spectral distribution correlations among neighboring pixels as well as correlations among spectral bands at each pixel. The proposed techniques overcome the shortcomings of the previous approach by joint modeling of spectral feature distributions among neighboring pixels, and by partially modeling correlations across spectral bands with a set of semi-tied covariance matrices in Gaussian mixture densities. The GMD parameters and semi-tied transformation matrices are first estimated by an iterative maximum likelihood estimation (MLE) algorithm of Expectation-Maximization (EM) [5], and the parameters are next tuned by a minimum classification error (MCE) training algorithm to enhance the discriminative power of the statistical classifiers.

Semi-tied covariance modeling has been proposed previously for acoustic modeling of speech signals [6,7,8]. In the semi-tied method, diagonal covariance matrices are clustered into groups, and for each group, a shared transformation function acts on the individual diagonal covariance matrices to produce individual full covariance matrices. The semi-tied covariance method offers a reasonable compromise between modeling accuracy and requirement of data for reliable parameter estimation. The MCE training technique has been extensively studied in the area of speech recognition [9,10]. The motivation behind MCE training is that MLE-based parameter estimation optimizes distribution fitting but not classification accuracy. In MCE training, classification errors are embedded into a loss function that is directly

Battlespace Digitization and Network-Centric Warfare II, Raja Suresh, William E. Roper, Editors, Proceedings of SPIE Vol. 4741 (2002) © 2002 SPIE · 0277-786X/02/$15.00

minimized through parameter estimation. As the result, the discriminative power of statistical classifiers can be significantly enhanced for highly confusable data classes.

This paper is organized as four sections. In Section 2, the statistical modeling and training algorithms are described. In Section 3, the experimental data, procedure, and results are discussed and summarized. A conclusion is made in Section 4.

2. Statistical Modeling Techniques

The remotely sensed imagery pixels are defined as the basic classification unit. The multispectral measurement data at each pixel t constitute a feature vector x_t. The dimension of x_t is denoted by D, and the total number of land cover classes is denoted by N. In the following, the context-independent GMD model of spectral feature vectors as proposed in [2] is first reviewed, and the context-dependent statistical models that are proposed in the current work are then described.

2.1. Context-independent modeling

The GMD probability density function (pdf) of x_t given the land cover class i is defined as

$$f\left(x_t|\lambda^{(i)}\right) = \sum_{q=1}^{M} \alpha_q^{(i)} \mathcal{N}\left(x_t|\mu_q^{(i)}, \Sigma_q^{(i)}\right) \tag{1}$$

where $\left(\mu_q^{(i)}, \Sigma_q^{(i)}\right)$ are the mean vector and covariance matrix of the q^{th} Gaussian density, $\alpha_q^{(i)}$'s are mixture weights satisfying the stochastic constraint $\alpha_q^{(i)} \geq 0,\ \sum_{q=1}^{M} \alpha_q^{(i)} = 1$, and M is referred to as mixture size. The covariance matrices are constrained to be diagonal, i.e., $\Sigma_q^{(i)} = \mathrm{dia_g}\left(\sigma_{\mathrm{q,1}}^{2(\mathrm{i})}, \sigma_{\mathrm{q,2}}^{2(\mathrm{i})}, \cdots, \sigma_{\mathrm{q,D}}^{2(\mathrm{i})}\right)$. Given a set of training data for class i, the parameters $\lambda^{(i)} = \left\{\alpha_q^{(i)}, \mu_q^{(i)}, \Sigma_q^{(i)}, q = 1, \cdots, M\right\}$ are estimated by the EM algorithm [5]. The maximum-likelihood (ML) based optimal decision rule for pixel t is defined as

$$s_t^* = \mathrm{ar_g} \max_{1 \leq i \leq N} f\left(x_t|\lambda^{(i)}\right) \tag{2}$$

.

2.2. Context-dependent modeling

The proposed context-dependent modeling is tied with a Bayesian decision rule, where the posterior probability of a pixel's class is based on spectral measurements in a 3x3 pixel block, with the pixel being the center of the block. In order to alleviate the sparse data problem in model training, several approximations were made in modeling the interaction of pixels within a block. The approximated models and the associated classifiers are described below.

2.2.1. Four directional context analysis

As shown in Fig. 1, a 3x3 pixel block is decomposed in four directions of $l = 1, 2, 3, 4$. Each direction consists of three pixels, including the center pixel t. Denote the joint prior probability of three pixels in the direction l by $P\left(s_t = i, s_{t,-1}^{(l)} = j, s_{t,1}^{(l)} = k\right) = \pi_{i,j,k}^{(l)}$. The posterior class probability of the pixel t given the spectral vectors in direction l is given by

$$\begin{aligned} & P\left(s_n = i|x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)}\right) \\ & = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)}|s_t = i, s_{t,-1}^{(l)} = j, s_{t,1}^{(l)} = k\right)\pi_{i,j,k}^{(l)}}{p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)}\right)} \\ & \triangleq \frac{\gamma_t^{(l)}(i)}{p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)}\right)} \end{aligned} \tag{3}$$

$s_{t,1}^{(4)}$	$s_{t,1}^{(3)}$	$s_{t,1}^{(2)}$
$s_{t,-1}^{(1)}$	s_t	$s_{t,1}^{(1)}$
$s_{t,-1}^{(2)}$	$s_{t,-1}^{(3)}$	$s_{t,-1}^{(4)}$

Figure 1. Four directional context analysis

where $\gamma_t^{(l)}(i)$ denotes the joint pdf of the class configuration (i,j,k) and the three spectral vectors in direction l. For maximum *a posteriori* (MAP) classification, the marginal pdf $p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)}\right)$ has no effect and therefore can be dropped. The decision on the land cover class of pixel t is defined to be based on $\gamma_t^{(l)}(i)$ in all four directions, i.e., $s_t^* = \arg\max_{1\le i\le N} \prod_{l=1}^{4} \gamma_t^{(l)}(i)$.

The prior probabilities of class configurations are estimated by normalized occurrence counts from the training set Ω_T. Define

$$\delta_t^{(l)}(i,j,k) = \begin{cases} 1 & if \ \ in\ direction\ l\ s_t = i\ and\ s_{t,-1} = j\ and\ s_{t,1} = k \\ 0 & otherwise \end{cases}$$

Then

$$\pi_{i,j,k}^{(l)} = \frac{\sum_{t\in\Omega_T} \delta_t^{(l)}(i,j,k)}{|\Omega_T|}$$

A direct modeling of the joint spectral distribution $p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)} | s_t = i, s_{t,-1}^{(l)} = j, s_{t,1}^{(l)} = k\right)$ is difficult due to rare occurrences of some class configurations related to small-sized classes. Two approximations to the probability distributions are made for coping with this problem.

A. Context-independent pdf

Given the land cover class configuration in each direction l, the spectral distributions of the three pixels in the direction are assumed independent, i.e.,

$$\begin{aligned} & p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)} | s_t = i, s_{t,-1}^{(l)} = j, s_{t,1}^{(l)} = k\right) \\ & \approx p(x_t | s_t = i) p\left(x_{t,-1}^{(l)} | s_{t,-1}^{(l)} = j\right) p\left(x_{t,1}^{(l)} | s_{t,1}^{(l)} = k\right) \end{aligned}$$

The pixel-based pdfs are taken to be Gaussian mixture densities, and the estimation of the pdf parameters is identical to that described in 2.1.

B. Context-dependent pdf

In each direction l, the joint spectral distribution for a given land-cover class configuration is approximated as a product of three conditional distributions:

$$\begin{aligned} & p\left(x_t, x_{t,-1}^{(l)}, x_{t,1}^{(l)} | s_t = i, s_{t,-1}^{(l)} = j, s_{t,1}^{(l)} = k\right) \\ & \approx p\left(x_t | s_t = i, s_{t,-1}^{(l)} = j, s_{t,1}^{(l)} = k\right) p\left(x_{t,-1}^{(l)} | s_t = i, s_{t,-1}^{(l)} = j\right) p\left(x_{t,1}^{(l)} | s_t = i, s_{t,1}^{(l)} = k\right) \end{aligned} \tag{4}$$

where the pdf of the center pixel depends on the class configuration of the three pixels, and those of the two boundary pixels depend on the classes of the center pixel and themselves. As such, context-dependent modeling is limited within the 3x3 block.

Each conditional pdf is defined as a Gaussian mixture density. In estimating parameters of the conditional pdfs, training data are partitioned into subsets according to the land cover class configurations as required in Eq.(4), for $l = 1, 2, 3, 4$. The training data of a subset are used to estimate the parameters of the corresponding conditional pdf by the EM algorithm.

2.2.2. Two neighborhood context analysis

As shown in Fig. 2, a 3x3 pixel block is decomposed around the center pixel into two neighborhoods of $l = 1, 2$. Each neighborhood has 4 pixels. Define the joint class probability of the center pixel and the neighborhood l as $P\left(s_t = i, s_{t,1}^{(l)} = j, s_{t,2}^{(l)} = k, s_{t,3}^{(l)} = m, s_{t,4}^{(l)} = n\right) = \pi_{i,j,k,m,n}^{(l)}$. Then the posterior probability of the center pixel's class given x_t and the spectral vectors in the neighborhood l is defined as

$$P\left(s_t = i | x_t, x_{t,1}^{(l)}, x_{t,2}^{(l)}, x_{t,3}^{(l)}, x_{t,4}^{(l)}\right)$$
$$= \frac{\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} p\left(x_t, x_{t,1}^{(l)}, x_{t,2}^{(l)}, x_{t,3}^{(l)}, x_{t,4}^{(l)} | s_t = i, s_{t,1}^{(l)} = j, s_{t,2}^{(l)} = k, s_{t,3}^{(l)} = m, s_{t,4}^{(l)} = n\right) \pi_{i,j,k,m,n}^{(l)}}{p\left(x_t, x_{t,1}^{(l)}, x_{t,2}^{(l)}, x_{t,3}^{(l)}, x_{t,4}^{(l)}\right)}$$

The prior probabilities of class configurations are again estimated by occurrence counts in Ω_T. The joint pdf of five spectral vectors is approximated in a similar way as made in four directional analysis.

For the case of context-independent pdf, the spectral distributions of the five pixels are assumed to be independent, i.e.,

$$p\left(x_t, x_{t,1}^{(l)}, x_{t,2}^{(l)}, x_{t,3}^{(l)}, x_{t,4}^{(l)} | s_t = i, s_{t,1}^{(l)} = j, s_{t,2}^{(l)} = k, s_{t,3}^{(l)} = m, s_{t,4}^{(l)} = n\right)$$
$$= p(x_t | s_t = i) p\left(x_{t,1}^{(l)} | s_{t,1}^{(l)} = j\right) p\left(x_{t,2}^{(l)} | s_{t,2}^{(l)} = k\right) p\left(x_{t,3}^{(l)} | s_{t,3}^{(l)} = m\right) p\left(x_{t,4}^{(l)} | s_{t,4}^{(l)} = n\right)$$

For the case of context-dependent pdf, the spectral distribution of the five pixels is approximated as a product of five conditional pdfs, i.e.,

$$p\left(x_t, x_{t,1}^{(l)}, x_{t,2}^{(l)}, x_{t,3}^{(l)}, x_{t,4}^{(l)} | s_t = i, s_{t,1}^{(l)} = j, s_{t,2}^{(l)} = k, s_{t,3}^{(l)} = m, s_{t,4}^{(l)} = n\right)$$
$$\approx p\left(x_t | s_t = i, s_{t,1}^{(l)} = j, s_{t,2}^{(l)} = k, s_{t,3}^{(l)} = m, s_{t,4}^{(l)} = n\right)$$
$$\times p\left(x_{t,1}^{(l)} | s_t = i, s_{t,1}^{(l)} = j\right) p\left(x_{t,2}^{(l)} | s_t = i, s_{t,2}^{(l)} = k\right) p\left(x_{t,3}^{(l)} | s_t = i, s_{t,3}^{(l)} = m\right) p\left(x_{t,4}^{(l)} | s_t = i, s_{t,4}^{(l)} = n\right)$$

Each conditional pdf is assumed to be a Gaussian mixture density, and the training on the GMD parameters follows the procedure outlined in section 2.2.1. B. It is worth noting that each pdf of the center pixel is conditioned on class configurations of five pixels and therefore the two-neighborhood analysis is N^2 folds more complex than the four–directional analysis. A direct consequence is that the size of training data subsets that correspond to five-pixel class configurations are in general much smaller than that of 3–pixel class configurations.

Denote the joint probability of class configuration and spectral vectors of the pixel t and its neighborhood l by $\gamma_t^{(l)}(i)$. The classification on pixel t is defined to be based on $\gamma_t^{(l)}(i),\ l = 1, 2$, i.e., $s_t^* = \arg\max_{1 \le i \le N} \prod_{l=1}^{2} \gamma_t^{(l)}(i)$.

$s_{t,2}^{(2)}$	$s_{t,2}^{(1)}$	$s_{t,1}^{(2)}$
$s_{t,3}^{(1)}$	s_t	$s_{t,1}^{(1)}$
$s_{t,3}^{(2)}$	$s_{t,4}^{(1)}$	$s_{t,4}^{(2)}$

Figure 2. Two neighborhood context analysis

2.3. Semi-tied covariance modeling

In the Gaussian mixture density models of land-cover spectral data as discussed in Sections 2.1–2.2, each component Gaussian density is defined to have a diagonal covariance matrix. Since spectral measurements in different bands of Landsat TM sensor are somewhat correlated, using full covariance matrices in component Gaussian densities is expected to be more accurate for characterizing distributions of land-cover spectral data. Compared with a full covariance structure, advantages of a diagonal structure are that more reliable parameter estimates can be obtained from a limited amount of training data, and less time is needed in likelihood score computations. Semi-tied covariance structure lies in between diagonal and full covariance structures, where a small number of full covariance matrices are shared by a large set of Gaussian densities, whilst each Gaussian density maintains its own diagonal covariance matrix. Among the semi-tied methods, the one proposed in [8] allows maximum likelihood estimation of covariance parameters and it is adopted for use in the current work.

Consider a Gaussian mixture density model of a land cover class. For simplicity of notations, the class index i as used in Eq(1) is dropped in the following discussions. In semi-tied covariance modeling, the covariance matrix Σ_q of the qth Gaussian density in a mixture consists of two elements, a diagonal covariance matrix $\Sigma_q^{(\mathrm{diag})}$ that is specific to q, and a nondiagonal matrix H_r that is shared by several Gaussian densities. The covariance matrix Σ_q is constructed from the two elements as

$$\Sigma_q = H_r \Sigma_q^{(\mathrm{diag})} H_r^T \tag{5}$$

where T denotes transpose.

For a given set of Gaussian densities, the structure of transformation tying needs to be specified, including the number of transformation matrices R, and the partition of Gaussian densities into subsets Ω_r, $r = 1, 2, \cdots, R$, where within Ω_r, the Gaussian densities share the transformation H_r. In [8], the tying structure was specified according to a hierarchical phonetic structure of speech, where each allophone was modeled by a diagonal Gaussian density, and allophones of the same monophone were tied to share a transformation matrix. Such a physical structure is not apparent in the land cover data, and therefore a data-driven procedure is developed to determine the tying structure within each Gaussian mixture density.

For a mixture density of size M, the feature dimension d^* that accounts for most significant separation of Gaussian densities is identified and used for grouping Gaussian densities into R clusters. For this purpose, a global mean vector of the mixture density is first computed as $\bar{\mu} = \sum_{q=1}^{M} \alpha_q \mu_q$, and the significant feature dimension d^* is then determined by $d^* = \arg\max_{1 \le d \le D} \sum_{q=1}^{M} \alpha_q (\mu_{q,d} - \bar{\mu}_d)^2$. In this dimension, the variances of Gaussian densities, $\left\{\sigma^2_{1,d^*}, \sigma^2_{2,d^*}, \cdots, \sigma^2_{M,d^*}\right\}$, are sorted and partitioned into R intervals on the real axis. The Gaussian densities that are grouped into the same interval r, denoted by $q \in \Omega_r$, are assigned to share the transformation H_r.

For Gaussian densities within Ω_r, a global covariance matrix Σ_r is first estimated, and an eigen-decomposition is then performed as $\Sigma_r = H_r \Sigma_r^{(\mathrm{diag})} H_r^T$. The eigen-vector matrix H_r is taken as the initial transformation matrix. Fixing the rest parameters of the Gaussian densities unchanged, the transformation matrix H_r and the diagonal covariance matrices $\Sigma_q^{(\mathrm{diag})}$, $q \in \Omega_r$, are then iteratively estimated by an EM-like algorithm. The estimation equations are similar to those derived in [8].

The semi-tied covariance structure allows better modeling of spectral correlations than a diagonal structure while avoiding large number of free parameters as in a full structure. At the classification stage, the complexity of likelihood score computation is only moderately increased over the diagonal structure. For a Gaussian density $q \in \Omega_r$, the likelihood of a spectral vector x_t can be computed as

$$\mathcal{N}(x_t; \mu_q, \Sigma_q) = \mathcal{N}\left(H_r^{-1} x_t; H_r^{-1}\mu_q, \Sigma_q^{(\mathrm{diag})}\right) |H_r|^{-1} \tag{6}$$

The transformation on the mean vector $H_r^{-1}\mu_q$ and the scaling by $|H_r|^{-1}$ can be performed prior to classification and therefore do not incur overhead. Extra computation comes from the transformation of spectral data $H_r^{-1} x_t$ for every

t and r. Due to sharing these matrices among Gaussian densities, the extra cost is much less than that required by a full covariance structure.

2.4. MCE training of GMDs

MCE training as formulated in the current work is based on the ML decision rule of Eq.(2), with the understanding that the model parameters include the semi-tied transformation matrices, i.e., $\lambda^{(i)} = \left\{\alpha_q^{(i)}, \mu_q^{(i)}, \Sigma_q^{(i)}, H_r^{(i)}, q = 1, \cdots, M,\ r = 1, \cdots, R\right\}$, and the Gaussian density likelihoods are evaluated by Eq.(6). This formulation of MCE does not match the approximate Bayesian decision rules defined for the context-dependent models. Nonetheless, the formulation is chosen for its simplicity. It is expected that the models tuned by MCEs should improve the classification accuracies of both ML and Bayesian classifiers.

Let the set of semi-tied GMD parameters of N classes be denoted by $\Lambda = \left\{\lambda^{(i)},\ i = 1, 2, \cdots, N\right\}$. Based on the ML decision rule, the discriminant function for classifying x_t to the land cover class i, denoted by C_i, is in effect the log likelihood of x_t given $\lambda^{(i)}$, i.e., $g_i(x_t; \Lambda) = \log f(x_t|\lambda^{(i)})$. If $x_t \in C_i$ but $g_i(x_t; \Lambda) < g_{i'}(x_t; \Lambda),\ i \neq i'$, then a misclassification occurs. Accordingly, a misclassification measure can be defined for class i as

$$d_i(x_t; \Lambda) = -g_i(x_t; \Lambda) + \log\left[\frac{1}{N-1}\sum_{i'=1, i' \neq 1,}^{N} \exp\left[g_{i'}(x_t; \Lambda)\eta\right]\right]^{1/\eta} \tag{7}$$

The misclassification measure is further embedded into a 0–1 loss function:

$$\mathcal{L}_i(x_t; \Lambda) = \frac{1}{1 + \exp\left(-\gamma d_i(x_t; \Lambda)\right)} \tag{8}$$

Correct classification corresponds to zero or small loss, and misclassification corresponds to large loss with an upper bound of 1. The shape of the sigmoid loss function varies with the parameter $\gamma > 0$: the larger the γ, the narrower the transition region of $\mathcal{L}_i(x_t; \Lambda) = 0$ to $\mathcal{L}_i(x_t; \Lambda) = 1$.

Given a set of training data $\{x_t,\ t = 1, 2, \cdots, T\}$, an empirical loss function on the training data set is defined as

$$\mathbf{L}(\Lambda) = \sum_{t=1}^{T}\sum_{i=1}^{N} \mathcal{L}_i(x_t; \Lambda) 1(x_t \in C_i) \tag{9}$$

where $1(A)$ is an indicator function that takes the value of 1 if the condition A is true and 0 otherwise. Minimizing the empirical loss is equivalent to minimizing total misclassification errors. The model parameters are therefore estimated by carrying out a gradient descent on $\mathbf{L}(\Lambda)$. In order to ensure that the estimated GMD weight parameters satisfy the stochastic constraints and the variance parameters remain positive, parameter transformations of $\alpha_q^{(i)} \rightarrow \tilde{\alpha}_q^{(i)}$ and $\sigma_{\mathrm{q,d}}^{(\mathrm{i})} \rightarrow \tilde{\sigma}_{\mathrm{q,d}}^{(\mathrm{i})}$ are performed as $\alpha_q^{(i)} = \frac{\exp\left(\tilde{\alpha}_q^{(i)}\right)}{\sum_{q'=1}^{M} \exp\left(\tilde{\alpha}_{q'}^{(i)}\right)}$ and $\tilde{\sigma}_{\mathrm{q,d}}^{(\mathrm{i})} = \log \sigma_{\mathrm{q,d}}^{(\mathrm{i})}$. In addition, $\mu_{\mathrm{q,d}}^{(\mathrm{i})} \rightarrow \tilde{\mu}_{\mathrm{q,d}}^{(\mathrm{i})}$ is performed as $\tilde{\mu}_{\mathrm{q,d}}^{(\mathrm{i})} = \frac{\mu_{\mathrm{q,d}}^{(\mathrm{i})}}{\sigma_{\mathrm{q,d}}^{(\mathrm{i})}}$. The set of transformed GMD parameters and the semi-tied transformation parameters is denoted by $\tilde{\Lambda}$, and parameter updates are performed with respect to $\tilde{\Lambda}$. A sequential mode is adopted for updating the model parameters from successive data samples x_t, and the training procedure goes multiple passes through the training sample set. Denote the training passes by $s = 1, 2, \cdots, S$. The sequential MCE training in the s^{th} pass is performed as

$$\tilde{\Lambda}_s(t+1) = \tilde{\Lambda}_s(t) - \epsilon_s \nabla_{\tilde{\Lambda}} \sum_{i=1}^{N} \mathcal{L}_i\left(x_t; \tilde{\Lambda}\right) 1(x_t \in C_i)|_{\tilde{\Lambda}=\tilde{\Lambda}_s(t)},\ \ t = 1, 2, \cdots, T \tag{10}$$

where ϵ_s controls the estimation step size. Within a training pass ϵ_s is fixed, and with training passes ϵ_s decreases to ensure convergence. Details of the gradient equations with respect to various model parameters are discussed in [11].

The initial model parameters $\Lambda_1(0)$ or $\tilde{\Lambda}_1(0)$ are taken as those trained by the EM algorithm. The model parameters obtained at the end of the s^{th} iteration, $\tilde{\Lambda}_s(T)$, are used as the initial model parameters in the $s+1^{th}$ iteration, $\tilde{\Lambda}_{s+1}(0)$. The training process terminates if the empirical loss converges or the number of iterations reaches a preset limit.

3. EXPERIMENTS

3.1. Experimental data

The Landsat Thermatic-Mapper (TM) scene 2534 (path 25 and row 34) was taken as the experimental data. The size of the scene is 7707×6867 pixels, and it covers central Missouri with a diversity of ecological regions and land cover classes. Six channels of the TM sensor at 30 m ground resolution were used as spectral features. The scene data included two seasonal images: spring (May 1992) and fall (September 1992). For each pixel of the Landsat TM scene, the six-channel spectral measurements of the two seasons were stacked together as a 12-dimensional spectral feature vector. Definition of the eight land cover classes and their pixel-count distributions in the scene 2534 are provided in Table 1.

Table 1 Definition and pixel-count histogram of eight land cover classes in 2534 scene.

No.	1	2	3	4	5	6	7	8
Class	Urban	Cropland	Shrubland	Barren	Vegetation	Forest	Woodland	Herbaceous
(%)	0.71	0.62	0.01	2.01	0.11	40.95	6.93	48.63

The ground truth data for the scene were provided by Missouri Resource Assessment Partnership (MoRAP) [12]. The class labels were pixel-based and were generated by combining an unsupervised spectral clustering with a human-expert supervised class labeling. The pixel-based land cover class labels in scene 2534 is referred to as the MoRAP map. Based on the assumption that the MoRAP land cover assignment was entirely correct, the pixel labels were used to train and evaluate the classifiers.

A fixed percentage of pixels from each class of the 2534 scene, i.e., 1%, was taken for training the statistical models. In the case of pixel-block based modeling, the percentage was counted by the center pixels of the blocks. The 2534 scene pixels excluding those used in the training set were sampled for use in the test set. An informal classification experiment showed that taking 1% of the 2534 scene data from each class was comparable in performance with taking larger fixed percentages of data. Therefore, in the following experiments, 1% data were sampled from each class in the 2534 scene for use as the test data.

3.2. Model training procedures

Each land cover class was modeled by a Gaussian mixture density. The size of a mixture density was made to be dependent on the training sample size of its class. Let N denote the number of training samples for a class. An empirical rule for choosing mixture size M was set as $M = 2^{\lfloor \log_{10} N \rfloor + 1}$. The resulting mixture sizes ranged from 2 through 32 from small to large classes.

MLE-based parameter estimation of a Gaussian mixture density model consisted of two steps: a vector-quantization (VQ) [13] based initialization and EM. In the initialization phase, M codewords were first produced by VQ as the Gaussian mean vectors. The training data samples were partitioned by the codewords, and one sample covariance matrix was estimated for each partition. Mixture weights were initialized by relative sizes of data samples in the M partitions. In the EM phase, 20 iterations were performed to refine the GMD parameters, where the estimation was basically converged.

The prior probabilities of land cover class configurations were estimated by normalized occurrence counts as discussed in Sections 2.2.1 and 2.2.2. These parameters were held fixed throughout the rest training procedures.

Semi-tied covariance modeling was applied to the context-independent GMDs of Sections 2.1 and 2.2. In estimation of the transformation matrices, the tying size was fixed as $|\Omega_r| = 4$, i.e., four Gaussian densities shared one transformation matrix. The number of iterations in covariance estimation was set as 5.

In MCE training, the parameter η of Eq.(7) was set as 1, the parameter γ of Eq. (8) was set as 0.3, the step-size parameter of Eq.(10) was set as $\epsilon_s = 0.01/s^{0.3}$, and number of passes was set as 5.

3.3. Results

Experimental evaluations were performed on the modeling techniques described in Sections 2.1 and 2.2, referred to as MLE-diagonal-GMD, on the semi-tied covariance modeling method of Section 2.3, referred to as MLE-semi-tied-GMD, and on the MCE training method of Section 2.4, referred to as MCE-semi-tied-GMD. The following abbreviations are used to denote the models described in Sections 2.1–2.2:

CI-GMD: context-independent GMD

4D-CI-GMD: 4-directional context analysis with context-independent GMD

4D-CD-GMD: 4-directional context analysis with context-dependent GMD

2N-CI-GMD: 2-neighborhood context analysis with context-independent GMD

2N-CD-GMD: 2-neighborhood context analysis with context-dependent GMD

Experimental results are summarized in classification accuracy, where the overall accuracy was computed by dividing the number of correctly classified pixels by the total number of pixel samples in the test set. The results of MLE-diagonal-GMD are shown in Table 2 for each model and each land cover class, and the results of MLE-diagonal-GMD, MLE-semi-tied-GMD and MCE-semi-tied-GMD are compared in Table 3 in terms of overall classification accuracy.

Table 2 Classification accuracy of MLE trained diagonal GMDs (%).

Classes	1	2	3	4	5	6	7	8	Overall
CI-GMD	67.23	62.37	36.11	90.33	36.20	78.92	64.82	76.93	76.92
4D-CI_GMD	66.13	51.93	2.38	92.25	27.96	87.13	47.53	86.30	83.64
4D-CN-GMD	68.43	34.05	0.00	92.32	40.05	89.55	13.52	90.28	85.87
2N-CI-GMD	55.72	31.62	0.00	91.91	27.34	88.38	23.48	87.23	83.97
2N-CD-GMD	12.13	5.49	0.00	86.34	3.56	92.17	11.59	93.67	86.65

Table 3 Comparison of baseline and proposed methods on overall classification accuracy (%).

Classes	MCE-diagonal-GMD	MLE-semi-tied-GMD	MCE-semi-tied-GMD
CI-GMD	76.92	78.97	81.85
4D-CI-GMD	83.64	84.36	85.54
2N-CI-GMD	83.97	84.69	85.88

The results verified that the proposed context-dependent models, the semi-tied covariance modeling, and the MCE training method successfully improved classification accuracy over the baseline model CI-GMD. On the other hand, it is observed that compared with CI-GMD, the context-analysis based models appeared to yield somewhat unbalanced results across land cover classes. This problem is attributed to the unbalanced training sample sizes of the land cover classes, where under context analysis, a significantly larger number of parameters need to be estimated for models in enumerated contexts, and the sparse data problem became accentuated for small classes. To compensate for the problem, the training sizes of the eight classes were adjusted as follows: 10% for classes 1, 2 and 4, 50% for classes 3 and 5, and 1% for classes 6, 7, and 8. The models were retrained for the case of 4D-CD-GMD. The classification accuracy of the eight land cover classes are shown in Table 4 and are seen to become much more balanced.

Table 4 Classification accuracy of 4D-CD-GMD with adjusted training sample sizes for small classes (%).

Classes	1	2	3	4	5	6	7	8	Overall
4D-CD-GMD	69.27	33.07	71.08	88.18	67.99	89.66	36.63	89.88	84.78

4. CONCLUSION

Novel techniques of statistical modeling and training are proposed and evaluated for improved classification on TM land cover data. Based on overall accuracy results from classification of eight land cover classes, the following conclusions can be made on the modeling and training techniques discussed in this paper: 1) the proposed approximate Bayesian decision rules that incorporate prior probabilities of class configurations are superior to MLE based noncontextual decision rule; 2) the proposed context-dependent GMDs are superior to context-independent GMDs in modeling multispectral data distributions; 3) the proposed semi-tied covariance modeling is superior to diagonal covariance modeling in capturing spectral correlations in pixel-based spectral vectors; 4) the proposed MCE-based model training is superior to MLE-based model training in optimizing discriminative power of statistical classifiers. A number of directions are worthy of further investigations, including robust estimation of prior probability parameters of class configurations, more detailed tying structure in semi-tied covariance modeling, formulation of MCE training base on Bayesian decision rules, selection of training data for different land cover classes, and classifier-level fusion from seasonal imageries.

ACKNOWLEDGMENT

This work was supported in part by the NASA Stennis Space Center under Remote Sensing Award (ICREST) NAG-13–99014.

REFERENCES

1. Y. J. Hung and P. H. Swain, "Bayesian contextual classification based on modified M-estimates and Markov random fields," *IEEE Trans. on Geosc. Remote Sensing*, vol. 34, no. 1, pp.67-75, Jan. 1996.
2. Y. Zhu, Y. Zhao, K. Palaniappan, X. Zhou, X. Zhuang, "Optimal Bayesian classifier for land cover classification using Landsat TM data," *Proc. IEEE International Geosc. and Remote Sensing Symposium*, vol. 1, pp. 447–450, Honolulu, Hawaii, July 2000.
3. K. Palaniappan, F. Zhu, X. Zhuang, Y. Zhao, and A. Blanchard, "Enhanced binary tree genetic algorithm for automatic land cover classification," *Proc. IEEE International Geosc. and Remote Sensing Symposium*, vol. 2, pp. 688–692, Honolulu, Hawaii, July 2000.
4. A. Lobo, "Image segmentation and discriminant analysis for the identification of land cover units in ecology," *IEEE Trans. on Geosc. Remote Sensing*, vol. 35, no. 5, pp.1136-1145, Sept. 1997.
5. A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. Royal Statistical Society*, B 39, no. 1, pp. 1–38, 1977.
6. A. Ljolje, "The importance of cepstral parameter correlations in speech recognition," *Comput. Speech Lang.*, vol. 8, pp. 223–232, 1994.
7. M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, pp. 249–264, 1996.
8. M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, June 1999.
9. B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–266, 1997.
10. S. Katagiri, B.-H. Juang and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proceedings of IEEE*, vol. 86. no. 11, pp. 2345-2372, 1998.
11. Y. Zhao and X. Zhou, "Statistical approaches for land cover classification," Technical Report, Department of CECS, University of Missouri, Jan. 2001.
12. R. Drobney, T. Haithcoat, and D. Diamond, Missouri GAP Analysis Project, Final Report, Missouri Cooperative Fish and Wildlife Research Unit, Dec. 1999.
13. R. M. Gray, "Vector quantization," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, April 1980.