Optimal Bayesian Classifier for Land Cover Classification Using Landsat TM Data

Yuanxin Zhu, Yunxin Zhao, Kannappan Palaniappan, Xiaobo Zhou, Xinhua Zhuang

Multimedia Communications and Visualization Laboratory
Department of Computer Engineering and Computer Science
University of Missouri-Columbia, MO 65211
{yxzhu, zhuang, palani, zhao}@cecs.missouri.edu

ABSTRACT

An optimal Bayesian classifier using mixture distribution class models with joint learning of loss and prior probability functions is proposed for automatic land cover classification. The probability distribution for each land cover class is more realistically modeled as a population of Gaussian mixture densities. A novel two-stage learning algorithm is proposed to learn the Gaussian mixture model parameters for each land cover class and the optimal Bayesian classifier that minimizes the loss due to misclassification. In the first stage, the Gaussian mixture model parameters for a given land cover class is learned using the Expectation-Maximization algorithm. The Minimum Description Length principle is used to automatically determine the number of Gaussian components required in the mixture model without overfitting. In the second stage, the loss functions and the *a priori* probabilities are jointly learned using a multiclass perceptron algorithm. Preliminary results indicate that modeling the multispectral, multitemporal remotely sensed radiance data for land cover using a Gaussian mixture model is superior to using unimodal Gaussian distributions. Higher classification accuracies for eight typical land cover categories over one full Landsat scene in central Missouri are demonstrated.

INTRODUCTION

Land cover classification from satellite remote sensing data has been an active area of research and development since the 1970's. Multispectral, multitemporal data have been used for both supervised and automatic land cover and land use classification at different scales. Remotely sensed data from a variety of spaceborne and airborne instruments have been used including Landsat MSS, Landsat TM, SPOT HRV, IRS, and NOAA AVHRR. Various learning approaches including Bayesian learning [4], artificial neural networks [1], [2], [3] and decision tree learning [5] have been applied to automatic land cover classification. In supervised studies, the input typically consists of spectral data with labeled classes (i.e. ground truth or expert knowledge), and the output consists of assigning all of the input data to (usually unique) land cover classes. In order to differentiate land cover at the species level, ancillary data, such as latitude, longitude, elevation, slope, aspect, soil type, landform, etc. is often used.

Bayesian learning has been widely used as a theoretically robust foundation for the classification of remotely sensed data. Due to the difficulty in learning the loss caused by misclassification, the maximum *a posteriori* (MAP) estimate is frequently used. In order to obtain the MAP estimator, it is necessary to model both class-conditional and prior probabilities. However, prior information is difficult to model or obtain, in which case, MAP estimation reduces to maximum likelihood (ML) estimation. The ML classifier relies on estimates of the mean vector and covariance matrix for each land cover class under the assumption that each land cover class can be modeled by a single multivariate Gaussian distribution. This approach provides satisfactory results in many cases, but failure to use prior information and the assumption of a single Gaussian distribution typically limits classification accuracy.

In this paper, we explore the use of an Optimal Bayesian Classifier (OBC) using a Gaussian mixture model (GMM) and estimation of the loss function for land cover classification. The losses caused by misclassification and the *a priori* probabilities are jointly learned by using a multiclass perceptron algorithm. A two-stage learning process is formulated to reduce the misclassification rate over the training set. The result of the first step is the number of components and maximum likelihood estimate of the Gaussian distribution parameters of all components for each land cover class. Gaussian Mixtures are suitable approximations for modeling complex distributions [10] like the land cover classes based on multispectral, multitemporal features.

Finding a ML estimate of the GMM parameters is a non-linear constrained optimization problem. The Expectation-Maximization (EM) algorithm provides a general approach to iterative computation of the ML parameters. In order to select the number of components in the mixture model, we use the Minimum Description Length (MDL) principle which minimizes the encoding length of the model parameters and of the ML estimate residuals. The number of model parameter is automatically decided using the MDL principle.

OPTIMAL BAYESIAN CLASSIFIER

Suppose there are N land cover classes, $c_1, c_2, \ldots, c_N$. For each class c_i, $i = 1,2,\cdots,N$, we collect a representative set of training data, $(\mathbf{x}_i^1, \mathbf{y}_i^1)$, $(\mathbf{x}_i^2, \mathbf{y}_i^2)$, $\ldots$, $(\mathbf{x}_i^{n_i}, \mathbf{y}_i^{n_i})$, where

0-7803-6359-0/00/$10.00 © 2000 IEEE

$\mathbf{x}_i^j \in \mathbf{R}^2$, denotes pixel locations, n_i is the number of training instance and $\mathbf{y}_i^j \in \mathbf{R}^p$, $j = 1,2,\cdots,n_i$, represents the associated p-dimensional spectral feature vectors. The problem is how to learn an optimal classifier which will label new instances with high confidence, given limited training data. Ancillary data or contextual information is not used in this paper.

The ML classifier is a parametric classifier that relies on the second-order statistics of a Gaussian pdf model for each class. The classifier assigns a label, $c(\mathbf{y})$, to a new instance, y, based on following discriminate function:

$$c(\mathbf{y}) = \arg\max_{1 \le i \le N} g(y|c_i) \tag{1}$$

where $g(y|c_i)$ is the pdf for class c_i in the form of $g(\mathbf{y}|c_i) = N(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix respectively. Using ML hypothesis, the mean vector and covariance matrix are estimated from the training data by the estimators, $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$.

$$\hat{\boldsymbol{\mu}}_i = \left(\sum_{j=1}^{n_i} \mathbf{y}_i^j\right) \Big/ n_i \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_i = \left(\sum_{j=1}^{n_i} (\mathbf{y}_i^j - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i^j - \hat{\boldsymbol{\mu}}_i)^T\right) \Big/ (n_i) \tag{3}$$

For OBC, the class label of a new instance $\mathbf{y}$, $c(\mathbf{y})$, is assigned so that the loss caused by misclassification is minimized, i.e.

$$c(\mathbf{y}) = \arg\min_{1 \le i \le N} \sum_{k=1}^{N} l_{ik} P(c_k|\mathbf{y}) \tag{4}$$

where l_{ik} is the loss caused by assigning $\mathbf{y}$ to class c_i but $\mathbf{y}$ actually comes from population of class c_k, and $P(c_k|\mathbf{y})$ is the posterior probability. Substituting Bayes' theorem in (4), we have

$$c(\mathbf{y}) = \arg\min_{1 \le i \le N} \sum_{k=1}^{N} l_{ik}\, P(\mathbf{y}|c_k) P(c_k) / P(\mathbf{y}) \tag{5}$$

where $P(c_k)$ is a prior probability of class c_k. Since $P(c_k)$ is unknown, we can learn it jointly with l_{ik}. By denoting $L_{ik} = l_{ik} P(c_k)$, the class conditional pdf $p(\mathbf{y}|c_k)$, then (5) simplifies to

$$c(\mathbf{y}) = \arg\min_{1 \le i \le N} \sum_{k=1}^{N} L_{ik}\, p(\mathbf{y}|c_k). \tag{6}$$

We need to learn the class-conditional pdf, $p(\mathbf{y}|c_k)$, and L_{ik}, for the OBC. Under the assumption that each land cover class is generated by a GMM, we employ EM algorithm together with MDL principle to recover ML parameters of $p(\mathbf{y}|c_k)$. The L_{ik}, generalized loss values, are learned using a multiclass perceptron algorithm.

EM Algorithm for Gaussian Mixture Models

The population of each class c_i is modeled as a mixture of L_i subpopulations, with each component being a Gaussian distribution,

$$p(\mathbf{y}|c_i) = \sum_{l=1}^{L_i} w_i^l N(\mathbf{y}; \boldsymbol{\mu}_i^l, \boldsymbol{\Sigma}_i^l) \tag{7}$$

where $\boldsymbol{\mu}_i^l, \boldsymbol{\Sigma}_i^l$, w_i^l are, respectively, the mean vector, covariance matrix, and the weights of lth Gaussian component. For each class c_i, the weights are constrained by

$$\sum_{l=1}^{Li} w_i^l = 1. \tag{8}$$

Denoting the parameters of GMM for the ith class as $\Lambda_i = \{ w_i^l, \boldsymbol{\mu}_i^l, \boldsymbol{\Sigma}_i^l, l = 1,2,\cdots,L_i \}$. The negative log-likelihood for the GMM with parameters Λ_i and sample data $\{\mathbf{y}_i^j : j = 1,2,\cdots,n_i\}$ is given by

$$-\log(L(\Lambda_i)) = -\sum_{j=1}^{n_i} \log p(y_i^j|c_i). \tag{9}$$

To find the ML estimate of Λ_i, we use the EM algorithm which is a general approach to iteratively compute ML estimates when the observations are incomplete [6]. The incompleteness arises from the fact that the original population of a given data is unknown. If the population ownership were known for each data point, finding an ML estimate of the parameters would simply be a matter of finding the sample mean and covariance matrix for each population of data, then determining the weight of each component by the fraction of points in the sample belonging to each population.

The EM algorithm proceeds iteratively in two steps: E-step and M-step. Using the n th estimate of the GMM parameters, the E-step solves for the n th ownership probabilities, $P_l(\mathbf{y}_i^j)^{(n)}$, which is the probability that the data point, $\mathbf{y}_i^j$, belongs to population generated by l th Gaussian of class c_i.

$$P_l(\mathbf{y}_i^j)^{(n)} = \frac{w_i^{l(n)} N(\mathbf{y}_i^j, \boldsymbol{\mu}_i^{l(n)}, \boldsymbol{\Sigma}_i^{l(n)})}{\sum_{l'=1}^{L_i} w_i^{l'(n)} N(\mathbf{y}_i^j, \boldsymbol{\mu}_i^{l'(n)}, \boldsymbol{\Sigma}_i^{l'(n)})} \tag{10}$$

The M-step updates the ML estimates of the GMM parameters using the previous estimate of ownership probabilities. The iterative formula for updating w_i^l, $\boldsymbol{\mu}_i^l$, and $\boldsymbol{\Sigma}_i^l$ are as follows:

$$w_i^{l(n+1)} = \frac{1}{n_i} \sum_{j=1}^{n_i} P_l(\mathbf{y}_i^j)^{(n)}, \; l = 1,2,\cdots,L_i \tag{11}$$

$$\boldsymbol{\mu}_i^{l(n+1)} = \left(\sum_{j=1}^{n_i} P_l(\mathbf{y}_i^j)^{(n)} \mathbf{y}_i^j\right) \Big/ \left(\sum_{j=1}^{n_i} P_l(\mathbf{y}_i^j)^{(n)}\right) \tag{12}$$

$$\boldsymbol{\Sigma}_i^{l(n+1)} = \frac{\sum_{j=1}^{n_i} P_l(\mathbf{y}_i^j)^{(n)} \left[(\mathbf{y}_i^j - \boldsymbol{\mu}_i^{l(n+1)})(\mathbf{y}_i^j - \boldsymbol{\mu}_i^{l(n+1)})^T\right]}{\sum_{j=1}^{n_i} P_l(\mathbf{y}_i^j)^{(n)}} \tag{13}$$

The EM algorithm generates a new estimate $\Lambda_i^{(n+1)}$ from an existing $\Lambda_i^{(n)}$, for $n = 0,1,\cdots$ until convergence. $\Lambda_i^{(0)}$ can be initialized using certain conventional clustering, say k-means clustering, with $w_i^{l(0)}$, $\boldsymbol{\mu}_i^{l(0)}$, $\boldsymbol{\Sigma}_i^{l(0)}$ computed as the sample estimates for each cluster l.

Minimum Description Length Encoding

The ML estimate of GMM parameters is unable to evaluate the complexity of the mixture model that is the number of Gaussian components for each land cover class. We address this problem by applying the MDL principle. The reason for choosing MDL is its information-theoretic grounding: the model that can be encoded most efficiently while explaining the observations is the best. For this purpose, the number of bits required to encode the model and the residuals is used. The goal then is to find the ML model parameters Λ_i that also minimize the total encoding length. The encoding has two parts, one part for the model and another for the data using the model. The overall code length for class c_i to be minimized is

$$C\left(\left\{\mathbf{y}_i^j\right\}\Lambda_i\right) = C_M(\Lambda_i) + C_D\left(\left\{\mathbf{y}_i^j\right\}|\Lambda_i\right) \qquad (14)$$

where C, C_M, and C_D denote the appropriate encoding length in terms of bits for the OBC estimator, model parameters, and data residuals respectively.

The model parameters consist of three different components: weights, means, and covariance matrices. For computing the coding cost of these real-valued parameters, the expression derived by Rissanen [7] in his optimal precision analysis is used. For encoding K independent real-valued parameters characterizing a distribution used to encode D data points, the code length is $(K/2)\log D$. Thus $C_M(\Lambda_i)=(K/2)\log n_i$, where K is the total number of parameters. Furthermore, we need to encode the data given the model. Since we know the likelihood of data from the mixture model, the optimal number of bits required to encode this is just the negative log-likelihood [7]. Therefore, this term is directly derived from the negative log-likelihood of the data given the model, presented in (9). Under the assumption of Gaussian distribution of the residual, and if the residuals are quantized to the nearest ε, their real precision, can be computed according to [8],

$$P(\mathbf{y}_i^j|\Lambda_i) \approx \varepsilon \sum_{l=1}^{L_i} w_i^l N\left(\mathbf{y}_i^j;\boldsymbol{\mu}_i^l,\boldsymbol{\Sigma}_i^l\right). \qquad (15)$$

Substituting (15) into (9) and eliminating the terms independent of L_i, the total encoding length is

$$C\left(\left\{\mathbf{y}_i^j\right\}\Lambda_i\right) = (K/2)\log n_i - \sum_{j=1}^{n_i}\log\left(\sum_{l=1}^{L_i} w_i^l N\left(\mathbf{y}_i^j;\boldsymbol{\mu}_i^l,\boldsymbol{\Sigma}_i^l\right)\right). \qquad (16)$$

Algorithm for Estimating GMM Parameters

Equation (16) is the expression for the complete encoding lengths of the models and the data given the models for class c_i. Ideally, optimization of these encoding lengths with respect to all the unknowns should be preformed. However, this is prohibitively expensive given the large parameter space. So we use a sequential approach by alternating steps of ML estimation of the parameters followed by evaluation of the model size using the MDL principle. We use an ascending greedy procedure which, given a lower bound on the number of models, incrementally computes the encoding length until it reaches a minimum description length. The first stage of the OBC consists of three different parts: the initialization step, the EM step, and the MDL step.

Multiclass Perceptron Algorithm

We learn L_{ik}'s (see (6)) using a perceptron algorithm [9], following is the derivation of a multiclass perceptron algorithm. By the definition of loss function, we have

$$\sum_{k=1}^{N} L_{ik} P(\mathbf{y}_i^j|c_k) < \sum_{k=1}^{N} L_{nk} P(\mathbf{y}_i^j|c_k) \qquad (17)$$
$$j = 1,2,\cdots n_i,\ n \neq i,\ n,i = 1,2,\cdots,N$$

Let $\mathbf{P}_i^j = \left[P(\mathbf{y}_i^j|c_1), P(\mathbf{y}_i^j|c_2),\cdots,P(\mathbf{y}_i^j|c_N)\right]^T$ and $\mathbf{L}_i = (L_{i1}, L_{i2},\cdots,L_{iN})^T$, then (18) can be written as

$$\mathbf{L}_i^T \cdot \mathbf{P}_i^j < \mathbf{L}_n^T \cdot \mathbf{P}_i^j \qquad (18)$$

Based on (18), the multiclass perceptron algorithm with parameter η, learning rate, is given by:

Initialize $\mathbf{L}_i^0$, $i = 1,2,\cdots,N$.
while not converged
  for $i \leftarrow 1$ *to* N
    for $n \leftarrow 1$ *to* N, *but* $n \neq i$
      for $j \leftarrow 1$ *to* n_i
        if $\mathbf{L}_i^{T(t)} \cdot \mathbf{P}_i^j \geq \mathbf{L}_n^{T(t)} \cdot \mathbf{P}_i^j$
          then $\mathbf{L}_i^{(t+1)} = \mathbf{L}_i^{(t)} - \eta \mathbf{P}_i^j$
          else $\mathbf{L}_n^{(t+1)} = \mathbf{L}_n^{(t)} + \eta \mathbf{P}_i^j$

EXPERIMENTAL RESULTS

We evaluate the proposed approach using a data set with supervised classification for the entire state of Missouri at 30m resolution and 8 land cover classes, shown in Table 1. The data provided by the Missouri Resource Assessment Partnership is referred to as the MoRAP Phase II Missouri Land Cover dataset. Our preliminary experiment is based on scene 2533 which covers central Missouri near Columbia with a diversity of ecological regions and land cover classes. The spectral features are from Landsat TM data at two scanning time (May and September 1992 respectively). This provides a total 12 spectral radiance measurements for each pixel (six TM channels at two times).

The results from four experiments are described. The data sets used in the first experiment was the ground truth data within scene 2533, a total of 2700 pixels, with 70% used for training and 30% for testing. The data set used in the

subsequent three experiments is the entire scene 2533 with 33,784,355 pixels of which different proportions were used for training and testing. Specifically, 5%, 10%, and 70% of the scene were used for training with the reminder 95%, 90%, and 30% used for testing, respectively. For the fourth experiment, the selected number of Gaussian components and the classification accuracy for each land cover class are shown in Table 1. The EM algorithm converges within 10 to 30 iterations in most cases.

The OBC results were compared to Simple Bayesian Classifier (SBC). The SBC models each land cover class using a single (unimodal) Gaussian distribution. Experimental results in Table 2 show that the classification performance of OBC is ~3% higher than that of SBC.

Table 1: Missouri Phase II land cover classes.

Class ID	Category description	# Gaussian components	Accuracy (%)
0	Urban	6	33.67
1	Cropland	5	77.39
2	Shrub land	4	58.23
3	Open water	3	99.34
4	Sparse vegetated	3	28.33
5	Forest	8	87.96
6	Woodland	7	62.60
7	Herbaceous	10	61.96

Table 2: Performance comparison between OBC and SBC.

Data Set	Accuracy (%) achieved by SBC		Accuracy (%) achieved by OBC	
	Training Set	Testing Set	Training Set	Testing Set
Ground truth data	81.22	82.35	84.19	84.57
Scene 2533 (5%)	73.57	73.53	77.63	76.68
Scene 2533 (10%)	73.50	73.55	77.24	76.88
Scene 2533 (70%)	73.23	74.00	77.01	76.95

SUMMARY

Our preliminary results show that modeling the multispectral, multitemporal remotely sensed radiance features for each land cover class using a GMM yields a better classifier than the single Gaussian model. However, to fully evaluate the performance of the proposed approach, additional experiments are needed to characterize convergence of the EM algorithm, suitability of the MDL for different classes, spectral heterogeneity of land cover classes across scenes, temporal stability of the classes, and generalization capability of the GMM. An independent accuracy assessment of the Phase II labeled land cover data set is needed to evaluate the quality of the training data. Comparison of the OBC approach with various decision tree approaches and neural network based classifiers is also being investigated [11]. Noise and outliers may have a significant influence on the performance of the EM estimator for the GMM parameters. Robust statistics methods can be used to improve the GMM estimation [12]

ACKNOWLEDGEMENTS This work was supported in part by NASA Stennis Space Center under Remote Sensing Award (ICREST) NAG-13-99014, NASA/GSFC Award NAG-5-3900, NASA MTPE Award NAG-5-6968, NAG-5-6283, and the NSF vBNS Award 9720668. The authors wish to thank David Diamond and MoRAP for providing the land cover and Landsat TM data set, Feng Zhu for data processing and analysis, and the ICREST group for discussions.

REFERENCES

[1] A. Baraldi and F. Parmiggiani, "A neural network for unsupervised categorization of multivalued input patterns: An application of satellite image clustering," *IEEE Trans. Geosci. Remote Sensing*, 33: 305-316, 1995.

[2] P. D. Heermann and N. Khazenie, "Classification of multispectral remote sensing data using a back-propagation neural network," *IEEE Trans. Geosci. Remote Sensing*, 30: 81-88, 1992.

[3] A. Strahler, D. Muchoney, J. Borak, et al., "MODIS land cover and land-cover change," MODIS land cover product algorithm theorectical basis document (version 5.0), May 1999.

[4] J. D. Paola and R. A. Schowengerdt, "A detailed comparison of back-propagation neural network and maximum-likelihood classifier for urban land use classification," *IEEE Trans. Geosci. Remote Sensing*, 33:981-996, 1995.

[5] X. Zhan, R. Defries, M. Hansen., J. Townshend, et al., "MODIS enhanced land cover and land cover change," MODIS land cover product algorithm theorectical basis document (version 2.0), April 1999.

[6] P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39(1), 1-38, 1977.

[7] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, 11(2): 416-431, 1983.

[8] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int'l J. of Comp. Vision*, 3(1): 73-102, 1989.

[9] M. Minsky and S. Papert, *Perceptrons*. Cambridge, MA: MIT Press, 1969.

[10] X. Zhuang, Y. Huang, K. Palaniappan, Y. Zhao, "Gaussian mixture density modeling, decomposition and applications", *IEEE Trans. Image Processing*, 5(9): 1293-1302, Sept. 1996.

[11] K. Palaniappan, F. Zhu, X. Zhuang, Y. Zhao, and A. Blanchard, "Enhanced binary tree genetic algorithm for automatic land cover classification", IEEE 2000 *Int. Geoscience and Remote Sensing Symposium*, Honolulu, Hawaii, July 24-28, 2000.

[12] X. Zhuang, K. Palaniappan, R. M. Haralick, "Highly robust statistical methods based on minimum-error Bayesian classification", In *Visual Information Representation, Communication and Image Processing*, Optical Engineering Series 64, Ed. C. W. Chen and Ya-Qin Zhang, Marcel-Dekker, New York, 1999, pp.415-430.