Robust Parameter Estimation For Mixture Model

Saldju Tadjudin
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
Phone (765) 494-9217
Fax (765) 494-3358
tadjudin@ecn.purdue.edu

David A. Landgrebe
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
Phone (765) 494-3486
Fax (765) 494-3358
landgreb@ecn.purdue.edu

Abstract -- An important problem in pattern recognition is the effect of limited training samples on classification performance. When the ratio of the number of training samples to the dimensionality is small, parameter estimates become highly variable, causing the deterioration of classification performance. This problem has become more prevalent in remote sensing with the emergence of a new generation of sensors. While the new sensor technology provides higher spectral and spatial resolution, enabling a greater number of spectrally separable classes to be identified, the needed labeled samples for designing the classifier remain difficult and expensive to acquire. Better parameter estimates can be obtained by exploiting a large number of unlabeled samples in addition to training samples using the expectation maximization (EM) algorithm under the mixture model. However, the estimation method is sensitive to the presence of statistical outliers. In remote sensing data, classes with few samples are difficult to identify and may constitute statistical outliers. Therefore, we propose a robust parameter estimation method for the mixture model. The proposed method assigns full weight to the sample from the main body of the data, but automatically gives reduced weight to statistical outliers. Experimental results show that the robust method prevents performance deterioration due to statistical outliers in the data as compared to the estimates obtained from EM approach.

INTRODUCTION

In a mixture model, data are assumed to consist of two or more distributions mixed in varying proportions. For remote sensing applications, it is a common practice to consider several "spectral subclasses" within each "information class" or ground cover type. Each of such spectral subclasses is considered to be normally distributed and classification is then performed with respect to the spectral subclasses. Under this model, we can regard remote sensing data as a mixture model fitted with normally distributed components. To estimate the model parameters in a mixture, a common approach is to apply the expectation maximization (EM) algorithm which is an iterative method for numerically approximating the maximum likelihood (ML) estimates of the parameters in a mixture model. Alternatively, it can be viewed as an estimation problem involving incomplete data in which each unlabeled observation is regarded as missing a label of its origin[1]. In [2] , the EM algorithm has been studied and applied to remote sensing data. It was shown that by assuming a mixture model and using both training samples and unlabeled samples in obtaining the estimates, the classification performance can be improved. Also, the Hughes phenomenon[3] can be delayed to a higher dimensionality and hence more features can be used to obtain better performance. In addition, the parameter estimates represent the true class distributions more accurately.

There are several factors affecting the convergence of the EM algorithm to the maximum likelihood estimates. First of all, the selection of training samples as initial estimates can affect the convergence to a great extent. In this work, the training set is assumed to provide a good initial estimate. Another factor that decides the performance of the EM algorithm is the presence of statistical outliers. Assume that the number of components have been decided and given by the training set. Statistical outliers are defined as those observations which have great discrepancy from the distributions of the mixture components. The EM algorithm assigns each observation to one of the components with the sample's posterior probability as its weight. Even though an outlying sample is inconsistent with distributions of all the defined components, it may still have a large posterior probability for one or more of the components. As a result, the iteration converges to erroneous solutions.

Unfortunately, for the analysis of remote sensing data, to arrive at a set of exhaustive classes is an iterative process by trial and error, and usually depends on the expertise of the user. In addition, there might be some scattered background pixels which are difficult or tedious to identify. These pixels form the so-called "information noise" which may constitute statistical outliers. Such outliers are usually eliminated using a chi-squared threshold[2] before applying the EM algorithm.

Work leading to this paper was supported in part by NASA under Grant NAG5-3975 and the Army Research Office under Grant DAAH04-96-1-0444.

In other words, pixels whose distances are greater than the threshold value are considered as outliers and are subsequently excluded from updating the estimates. However, a suitable threshold value is often difficult to select and is usually arbitrary. Consequently, "useful" pixels might be rejected as statistical outliers. In particular, as dimensionality increases, most pixels might be considered as outliers.

In this work, we propose a robust method to estimate the mean vector and covariance matrix for classifying remote sensing data under the mixture model. This approach assigns full weight to the training samples, but automatically gives reduced weight to unlabeled samples. Therefore, it avoids the risk of rejecting useful pixels while still limiting the influence of outliers in obtaining the ML estimates of the parameters. The experimental results show that the proposed robust method is effective in reducing the effect of statistical outliers as compared to the EM approach.

ROBUST ESTIMATION

The EM algorithm first estimates the posterior probabilities of each sample belonging to each of the component distributions, and then computes the parameter estimates using these posterior probabilities as weights. With this approach, each sample is assumed to come from one of the component distributions, even though it may greatly differ from all components. The robust estimation attempts to circumvent this problem by including the typicality of a sample with respect to the component densities in updating the estimates in the EM algorithm.

To incorporate a measure of typicality in the parameter estimation of the mixture density, the component densities $f_i(x|\mu_i,\Sigma_i)$ for x ^{p} are assumed to be a member of the family of p-dimensional elliptically symmetric densities with mean vector μ_i and covariance matrix Σ_i [4]:

$$|\Sigma_i|^{-1/2} f_S\{\delta_i(x;\mu_i,\Sigma_i)\}$$

where $\delta_i^2 = (x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)$. Typically, $f_S(\delta_i)$ is assumed to be the exponential of some symmetric function $\rho(\delta_i)$:

$$f_S(\delta_i) = \exp\{-\rho(\delta_i)\}.$$

Then, the likelihood parameter estimation for these component densities can be obtained by applying the expectation and maximization steps. Denoting the current and future parameter values by the superscripts "c" and "+", the iterative equations are derived as[4]:

$$\alpha_i^+ = \sum_{j=1}^{n} \tau_{ij}^c / n$$

$$\mu_i^+ = \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^c x_j \Big/ \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^c$$

$$\Sigma_i^+ = \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^+ (x_j-\mu_i^+)(x_j-\mu_i^+)^T \Big/ \sum_{j=1}^{n} \tau_{ij}^c$$

where $w_{ij} = \psi(\delta_{ij})/\delta_{ij}$ is the weight function and $\psi(\delta_{ij}) = \rho'(\delta_{ij})$ is the first derivative of $\rho(\delta_{ij})$. To limit the influence of large atypical samples, the covariance estimator is modified to be:

$$\Sigma_i^+ = \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^{+2} (x_j-\mu_i^+)(x_j-\mu_i^+)^T \Big/ \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^{+2}.$$

The weight function has been chosen to be $\psi(s)/s$ where $s = \delta_{ij}$. A popular choice of $\psi(s)$ is the Huber's ψ-function which is defined by $\psi(s) = -\psi(-s)$ where for $s > 0$

$$\psi(s) = \begin{cases} s & 0 \le s \le k_1(p) \\ k_1(p) & s > k_1(p) \end{cases}$$

for an appropriate choice of the "tuning" constant $k_1(p)$, which is a function of the dimensionality p. This selection of $\psi(s)$ gives:

$$\rho(s) = \begin{cases} \frac{1}{2}s^2 & 0 \le s \le k_1(p) \\ k_1(p)s - \frac{1}{2}k_1^2(p) & s > k_1(p) \end{cases}.$$

The value of the tuning constant is a function of dimensionality. It also depends on the amount of contamination in the data which is usually not known. Since the training samples are representative of the classes, it is desirable that they are given more emphasis in the updates of the estimates. Therefore, in the proposed approach, the training samples are assigned unit weight. To do so, the value of $k_1(p)$ is defined to be

$$k_1(p) = \max(\hat{d}_{ij})$$

where $\hat{d}_{ij}^2 = (y_{ij}-\mu_i)^T \Sigma_i^{-1}(y_{ij}-\mu_i)$ and y_{ij} is the training sample j from class i. In other words, the tuning constant is selected such that the training samples are given unit weight and the weights for the unlabeled samples are inversely proportional to the square root of their distances to the class

mean. Therefore, the weight assigned to each sample can be expressed as:

$$w_{ij} = \begin{cases} 1 & d_{ij} \le \max\left(\hat{d}_{ij}\right) \\ \max\left(\hat{d}_{ij}\right)/d_{ij} & \max\left(\hat{d}_{ij}\right) < d_{ij} < \infty \end{cases}$$

where $d_{ij}^2 = \left(x_j - \mu_i\right)^T \Sigma_i^{-1}\left(x_j - \mu_i\right)$ is the squared distance of unlabeled samples x_j. The iterative equations for the mean and covariance estimates can then be expressed as:

$$\mu_i^+ = \frac{\sum_{j=1}^{m_i} y_{ij} + \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^c x_j}{m_i + \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^c}$$

$$\Sigma_i^+ = \frac{\sum_{j=1}^{m_i}\left(y_{ij} - \mu_i^+\right)\left(y_{ij} - \mu_i^+\right)^T + \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^{+2}\left(x_j - \mu_i^+\right)\left(x_j - \mu_i^+\right)^T}{m_i + \sum_{j=1}^{n} \tau_{ij}^c w_{ij}^{+2}}.$$

EXPERIMENTAL RESULTS

In the following experiment we compare the performance of quadratic classifiers using the parameters estimated from training samples alone (ML), the EM algorithm (EM) and the proposed robust algorithm (REM). The data set consists of a portion of an AVIRIS data taken in June, 1992, which covers a mixture of agricultural/forestry land in the Indian Pine Test Site in Indiana. The water absorption bands were removed to leave a total of 200 bands. The data set is composed of 4 classes. The classes and the number of labeled samples are shown in Table 1.

The first experiment is intended to compare EM and REM without outliers in the data. To obtain data without outliers, we generate synthetic data using the statistics computed from the labeled samples of the four classes. Due to the limited labeled samples, we choose around 200 training samples and set the spectral channels at 10, 20, 50, 67 and 100. These channels are selected by sampling the spectral range at fixed interval. A total of 400 test samples are used for each class. Since the training samples are randomly selected from the sample pool, we perform 10 trials for each experiment and obtain the mean accuracy. The mean accuracy is shown in Fig. 1. The second experiment is conducted using the real data. The training samples are non-random and are selected by visual inspection. The result is shown in Fig. 2.

DISCUSSION AND CONCLUSION

Using synthetic data, the experimental result shows that when no outliers are present in the data, EM and REM have similar performance and both achieve better accuracy than ML without using additional unlabeled samples. Using the real image, REM performs better than ML and EM. This demonstrates that the scene contains outlying pixels which are not represented in the training set. In both experiments, the performance declines at 100 dimensions due to the Hughes effect. Specifically, it implies that 600 samples are inadequate to characterize 100 dimensional Gaussian distribution. In conclusion, the proposed robust method is effective in reducing the effect of outliers as compared to the EM algorithm. Further details of this algorithm can be found in [5].

Table 1 Class description of the AVIRIS data set

Class No.	Class Name	No. of Samples
1	Corn-notill	910
2	Soybeans-notill	638
3	Soybeans-min	1421
4	Grass	618

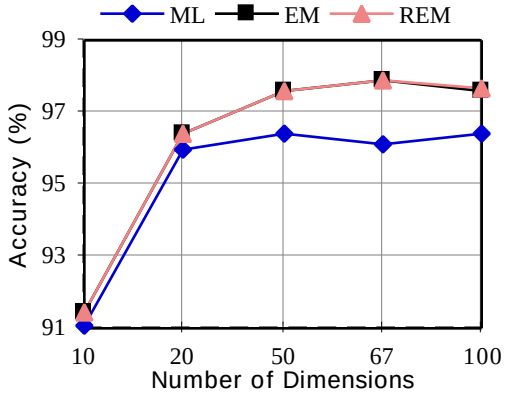

Fig. 1 Mean Accuracy Using Synthetic Data Without Outliers

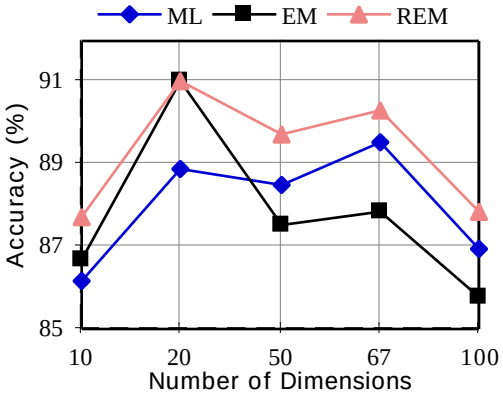

Fig. 2 Classification Accuracy Using Real Data

REFERENCES

[1] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood estimation from incomplete data via EM algorithm," J. R. Statist. Soc., Vol B39, pp. 1-38, 1977.
[2] B. Shahshahani and D.A. Landgrebe, Classification of Multi-spectral Data by Joint Supervised-Unsupervised

Learning, Purdue University, West Lafayette, IN, TR-EE 94-1, January 1994.
[3] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Inform. Theory., Vol IT-14, pp. 55-63, 1968.
[4] N.A. Cambell, "Mixture models and atypical values," Math. Geol., Vol 16, pp. 465-477, 1984.
[5] S. Tadjudin, Classification of High Dimensional Data with Limited Training Samples, Ph.D. Thesis, School of Electrical and Computer Engineering, Purdue University, 1998.