

MULTISOURCE DATA CLASSIFICATION USING A HYBRID SEMI-SUPERVISED LEARNING SCHEME

Ranga Raju Vatsavai*, Budhendra Badhuri

Shashi Shekhar†, Thomas E. Burk‡

Geographic Information Science and Technology
Computational Sciences and Engineering Division
Oak Ridge National Laboratory, Oak Ridge, TN 37831.

†Department of Computer Science
‡Remote Sensing Laboratory
University of Minnesota

ABSTRACT

In many practical situations thematic classes can not be discriminated by spectral measurements alone. Often one needs additional features such as population density, road density, wetlands, elevation, soil types, etc. which are discrete attributes. On the other hand remote sensing image features are continuous attributes. Finding a suitable statistical model and estimation of parameters is a challenging task in multisource (e.g., discrete and continuous attributes) data classification. In this paper we present a semi-supervised learning method by assuming that the samples were generated by a mixture model, where each component could be either a continuous or discrete distribution. Overall classification accuracy of the proposed method is improved by 12% in our initial experiments.

Index Terms— Semi-supervised learning, expectation maximization, GMM, multisource data

1. INTRODUCTION

A common task in analyzing remote sensing imagery is supervised classification, where the objective is to construct a classifier based on few labeled training samples and then to assign a label (e.g., forest, water, urban) to each pixel (vector, whose elements are spectral measurements) in the entire image. The commonly used maximum likelihood classifier (MLC) has two well known limitations. First, it works well if the land cover classes are spectrally separable. In reality, the classes under investigation are often spectrally overlapping as the reflectance recorded by remote sensing satellites for many of these thematic classes is dependent on several extraneous factors like terrain, soil type, moisture content, acquisition time, atmospheric conditions, etc. The usefulness of ancillary data for improving classification accuracy is well known, but there is no convenient multivariate statistical tool for modeling this multi-source data (i.e., images and ancillary geo-spatial data together). Previous studies [1], [2] have

focused on incorporating ancillary information into the MLC (typically via *a priori* term).

Second, MLC uses maximum likelihood estimation (MLE) technique for estimating class probability distribution parameters which requires large amounts of accurate training data. Collecting ground truth data for large number of samples is very difficult. Apart from time and cost considerations, in many emergency situations like forest fires, land slides, floods, it is impossible to collect accurate training samples. As a result, supervised learning is often carried out with small number of training samples, which leads to large variance in parameter estimates and thus higher classification error rates. Several approaches can be also be found in the literature that specifically deal with small sample size problems in supervised learning [3, 4, 5, 6, 7]. These methods are aimed at designing appropriate classifiers, feature selection, and parameter estimation so that classification error rates can be minimized while working with small sample sizes. However, only recently attempts have been made to incorporate unlabeled samples in supervised learning, which gave raise to new breed of techniques, collectively known as semi-supervised learning methods. Well-known studies in this area include, but not limited to [8, 9, 10, 11, 12]. The common thread between many of these methods is the Expectation Maximization (EM) [13] algorithm. Many of the semi-supervised learning methods pose class labels as the missing data and use the EM algorithm to improve initial (either guessed or estimated from small labeled samples) parameter estimates. Though previous studies [8, 14] showed that adding unlabeled training samples improves overall classification accuracy, little attention was given to extending semi-supervised learning for multisource data classification.

2. OUR APPROACH

In this paper, we provide a new hybrid semi-supervised learning method based on a mixture of discrete and continuous distributions. In typical semi-supervised approach, the population is assumed to be generated by a mixture of multivariate normal distributions for continuous attributes (e.g., re-

*Contact Author (vatsavairr@ornl.gov)

remote sensing images), or mixture of multinomial distributions for categorical attributes (e.g., text documents, ancillary geospatial data such as soil types, upland and lowlands). We now briefly describe semi-supervised learning in the following section, more details can be found in [14].

2.1. Semi-supervised Learning

First let us assume that each sample x_j comes from a super-population D , which is a mixture of a finite number (M) of populations D_1, \dots, D_M in some proportions $\alpha_1, \dots, \alpha_M$, respectively, where $\sum_{i=1}^M \alpha_i = 1$ and $\alpha_i \geq 0 (i = 1, \dots, M)$. Now we can model the data $D = \{x_i\}_{i=1}^n$ as being generated independently from the following mixture density.

$$p(x_i|\Theta) = \sum_{j=1}^M \alpha_j p_j(x_i|\theta_j) \quad (1)$$

Here $p_j(x_i|\theta_j)$ is the pdf corresponding to the mixture j and parameterized by θ_j , and $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$ denotes all unknown parameters associated with the M -component mixture density. For a multivariate normal distribution, θ_j consists of elements of the mean vectors μ_j and the distinct components of the covariance matrix Σ_j . The *log-likelihood* function for this mixture density can be defined as:

$$L(\Theta) = \sum_{i=1}^n \ln \left[\sum_{j=1}^M \alpha_j p_j(x_i|\theta_j) \right]. \quad (2)$$

In general, Equation 2 is difficult to optimize because it contains the \ln of a sum term. However, this equation greatly simplifies in the presence of unobserved (or incomplete) samples. Let us now pose X as an incomplete dataset, and assume that we have unobserved data $Y = \{y_i\}_{i=1}^n$ such that y_i tells us which component density generated each x_i . Assuming that we know the values of Y , the *log-likelihood* in Equation 2 can be simplified as:

$$L(\Theta) = \ln(P(X, Y|\Theta)) = \sum_{i=1}^n \ln(\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})). \quad (3)$$

However, in many supervised learning situations, the class labels (y_i)'s are not available. However, assuming the the initial parameters Θ^k can be guessed (as in clustering), or can be estimated (as in semi-supervised learning), we can easily compute $p_j(x_i|\theta_j^k)$ in eq. 1. Now, using Bayes' rule, we can compute

$$p(y_i|x_i, \Theta^k) = \frac{\alpha_{y_i}^k p_{y_i}(x_i|\theta_{y_i}^k)}{p(x_i|\Theta^k)} = \frac{\alpha_{y_i}^k p_{y_i}(x_i|\theta_{y_i}^k)}{\sum_{j=1}^M \alpha_j^k p_j(x_i|\theta_j^k)} \quad (4)$$

So, the expectation maximization (EM) algorithm at the first step maximizes the expectation of the *log-likelihood*

function, using the current estimate of the parameters and conditioned upon the observed samples. In the second step of the EM algorithm, called maximization, the new estimates of the parameters are computed. The EM algorithm iterates over these two steps until the convergence is reached. These two steps are formalized below:

E-step: At the i^{th} step of the iteration, where $\Theta^{(i-1)}$ is available, compute the expected value of

$$Q(\Theta, \Theta^{(k-1)}) = E \left[\ln p(X, Y|\Theta) | X, \Theta^{(k-1)} \right]. \quad (5)$$

This step is called the *expectation step*. In the function $Q(\Theta, \Theta^{(k-1)})$, the first argument Θ corresponds to the parameters that needs to be optimized by maximizing the *log-likelihood*, and the second argument $\Theta^{(k-1)}$ corresponds to the current estimate of the parameters that we used to evaluate the expectation.

M-step: Compute the new estimates of Θ by maximizing the $Q(\Theta, \Theta^{(k-1)})$, that is, find:

$$\theta^{(k)} = \arg \max_{\Theta} Q(\Theta, \Theta^{(k-1)}). \quad (6)$$

This second step is called the *maximization step*.

These two steps are repeated until convergence is reached. The *log-likelihood* function is guaranteed to increase until a maximum (local or global or saddle point) is reached. For multivariate normal distribution, the expectation $E[\cdot]$ (in Equation 6), which is denoted by p_{ij} , is the probability that Gaussian mixture j generated the data point i , and is given by:

$$p_{ij} = \frac{|\hat{\Sigma}_j|^{-1/2} e^{\{-\frac{1}{2}(x_i - \hat{\mu}_j)^t \hat{\Sigma}_j^{-1} (x_i - \hat{\mu}_j)\}}}{\sum_{l=1}^M |\hat{\Sigma}_l|^{-1/2} e^{\{-\frac{1}{2}(x_i - \hat{\mu}_l)^t \hat{\Sigma}_l^{-1} (x_i - \hat{\mu}_l)\}}} \quad (7)$$

The new estimates (at the k^{th} iteration) of parameters in terms of the old parameters at the M-step are given by the following equations:

$$\hat{\alpha}_j^k = \frac{1}{n} \sum_{i=1}^n p_{ij} \quad (8)$$

$$\hat{\mu}_j^k = \frac{\sum_{i=1}^n x_i p_{ij}}{\sum_{i=1}^n p_{ij}} \quad (9)$$

$$\hat{\Sigma}_j^k = \frac{\sum_{i=1}^n p_{ij} (x_i - \hat{\mu}_j^k)(x_i - \hat{\mu}_j^k)^t}{\sum_{i=1}^n p_{ij}} \quad (10)$$

2.2. Semi-supervised Learning for Multisource Data Classification

As explained previously, multisource data is a mixture of both continuous and discrete distributions. Let us now divide our attribute into two partitions: one consisting of all continuous variables and the other consisting of all discrete variables. We can now rewrite our mixture model given in Eq. 1 as following:

$$p(x_i|\Theta) = \sum_{j=1}^M \alpha_j \prod_{l=1}^2 p_{jl}(x_{il}|\theta_{jl}) \quad (11)$$

where θ_{jl} consists of the parameters of the distribution p_{jl} for the partition l . To reduce the complexity we used a knowledge based approach to stratify the geographic region into three broad categories, viz., uplands, lowlands and developed area. The main objective for this stratification is to split the geographic region into different spatial units where each spatial unit contains classes that are easily discriminable. So the discrete variables partition consists of a single attribute with three possible values. Finally we used expectation maximization algorithm to estimate the model parameters, where our model consists of six continuous attributes (corresponding to six channels in the ETM image) and a categorical attribute (generated using a knowledge based classification algorithm). We have conducted several experiments to evaluate the usefulness of our method in thematic classification of multisource geospatial datasets. We will present detailed description of the algorithm in full paper. We now briefly present the results in the following section.

3. EXPERIMENTAL RESULTS

We used a spring Landsat 7 scene, taken on May 31, 2000 over Cloquet town located in Carlton County, Minnesota. We designed two different experiments to validate our hypothesis that adding ancillary geospatial datasets and unlabeled training samples improve the classification performance.

We have used the following ancillary information: normalized density vegetation index (NDVI) and Tasseled Cap (images), transportation data (lines), National Wetland Inventory (NWI) data (polygons) and population data (polygons/attributes). We used a knowledge based approach [15] to generate a stratified image consisting of upland, lowland, and developed regions. This stratified image is used as a categorical attribute in our multisource classification experiment.

The labeled training data consists of 14 plots (2 plots per class), and unlabeled training data consists of 50 plots. For both of these experiments the test dataset was fixed and consisted of 205 plots. We trained three classifiers: MLC, Semi-supervised Classifier (SSL), and Multisource Semi-supervised classifier (SSL-MS). The estimates obtained by maximum likelihood and semi-supervised approaches (using

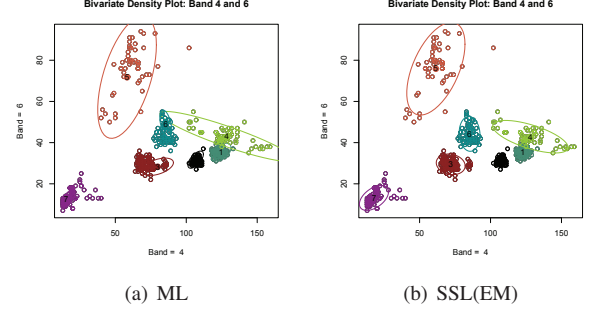


Fig. 1. Parameters Estimated using (a) ML, (b) SSL(EM).

C-ID	Class	MLC	SSL	SSL-MS
1	Hardwood.1	56.04	48.31	36.23
2	Hardwood.2	78.21	65.36	61.87
3	Conifer	43.30	80.84	86.59
4	Agriculture	84.19	94.02	98.72
5	Urban	100.00	91.11	97.78
6	Wetlands	6.50	28.38	66.50
7	Water	53.70	90.74	96.30
O	Overall	48.62	58.92	70.51

Fig. 2. Class and Overall Accuracy

expectation maximization) are summarized (in the form of bivariate density plots) in Figure 1. The individual class accuracy and overall classification accuracies were summarized in Figure 2. This figure (table) shows the great potential of our proposed classification scheme in small sample and multisource data classification problems. The plain semi-supervised learning method improved classification accuracy by 10% and on the other hand semi-supervised learning scheme on multisource data has resulted in improvement of about 22% over maximum likelihood classification.

4. CONCLUSIONS

In this study we presented a semi-supervised learning scheme for multisource data classification. This new scheme addresses two major limitations of the most widely used maximum likelihood classifier: small training samples and multisource data. Finite mixture modeling offers great flexibility in modeling multisource data. Initial experimental results showed an improvement of more than 20% as compared to MLC with training data of just 2 plots for class. Processing ancillary data to come up with meaningful stratified units is still an open problem. Further research is needed to automatically discover the stratified units from ancillary data for a given classification task. More experiments are needed to see the performance of the proposed algorithm in different geographic settings.

5. ACKNOWLEDGMENTS

We would like to thank our former collaborators Jamie Smedsmo, Ryan Kirk and Tim Mack at the Remote Sensing Laboratory of the University of Minnesota for useful comments and inputs into this research. The comments of Eddie Bright, Phil Coleman, and Veeraraghavan Vijayraj, have greatly improved the technical accuracy and readability of this paper.

Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725.

6. REFERENCES

- [1] A.H. Strahler, "The use of prior probabilities in maximum likelihood classification of remote sensing data," *Remote Sensing of Environment*, vol. 10, pp. 135–163, 1980.
- [2] F. Maselli, C. Conese, L. Petkov, and R. Resti, "Inclusion of prior probabilities derived from a non-parametric process into the maximum likelihood classifier," *Photogrammetric Engineering & Remote Sensing*, vol. 58, no. 2, pp. 201–207, 1992.
- [3] K. Fukunaga and Raymond R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, 1989.
- [4] Sarunas J. Raudys and Anil K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, 1991.
- [5] M. Skurichina and R. Duin, "Stabilizing classifiers for very small sample sizes," in *Proc. 10th Int. Conference on Pattern Recognition*, IEEE Computer Society Press, 1996, pp. 891–896.
- [6] S. Tadjudin and David A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosciences and Remote Sensing.*, vol. 37, no. 4, pp. 2113–2118, 1999.
- [7] R. Duin, "Classifiers in almost empty spaces," in *Proc. 15th Int. Conference on Pattern Recognition (Barcelona, Spain, Sep.3-7)*, vol. 2, IEEE Computer Society Press, 2000, pp. 1–7.
- [8] B. Shahshahani and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 32, no. 5, 1994.
- [9] T. Mitchell, "The role of unlabeled data in supervised learning," in *Proceedings of the Sixth International Colloquium on Cognitive Science, San Sebastian, Spain.*, 1999.
- [10] Sally Goldman and Yan Zhou, "Enhancing supervised learning with unlabeled data," in *Proc. 17th International Conf. on Machine Learning*. 2000, pp. 327–334, Morgan Kaufmann, San Francisco, CA.
- [11] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 2000.
- [12] F.G. Cozman, I. Cohen, and M.C. Cirelo, "Semi-supervised learning of mixture models," in *Twentieth International Conference on Machine Learning (ICML)*, 2003.
- [13] A.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*, vol. 39, no. 1, pp. 1–38, 1977.
- [14] Ranga Raju Vatsavai, Shashi Shekhar, and Thomas E. Burk, "A semi-supervised learning method for remote sensing data mining," in *ICTAI*, 2005, pp. 207–211.
- [15] Ranga R. Vatsavai, Thomas E. Burk, Paul V. Bolstad, Marvin E. Bauer, Sonja K. Hansen, Tim Mack, Jamie Smedsmo, and Shashi Shekhar, "Multi-spectral image classification using spectral and spatial knowledge," in *CISST*, 2001.