A Semi-supervised Learning Method for Remote Sensing Data Mining

Ranga Raju Vatsavai[1,2], Shashi Shekhar[1], and Thomas E. Burk[2]
[1]Department of Computer Science and Engineering, University of Minnesota
EE/CS 4-192, 200 Union Street. SE., Minneapolis, MN 55455. $[vatsavai|shekhar]$@cs.umn.edu
[2]Remote Sensing Laboratory, Dept. of Forest Resources, University of Minnesota
115, Green Hall, 1530 N. Cleveland Ave, St. Paul 55108. $[vrraju|tburk]$@gis.umn.edu

Abstract

New approaches are needed to extract useful patterns from increasingly large multi-spectral remote sensing image databases in order to understand global climatic changes, vegetation dynamics, ocean processes, etc. Supervised learning, which is often used in land cover (thematic) classification of remote sensing imagery, requires large amounts of accurate training data. However, in many situations it is very difficult to collect labels for all training samples. In this paper we explore methods that utilize unlabeled samples in supervised learning for thematic information extraction from remote sensing imagery. Our objectives are to understand the impact of parameter estimation with small learning samples on classification accuracy, and to augment the parameter estimation with unlabeled training samples to improve land cover predictions.

We have developed a semi-supervised learning method based on the Expectation-Maximization (EM) algorithm, and maximum likelihood and maximum a posteriori classifiers. This scheme utilizes a small set of labeled and a large number of unlabeled training samples. We have conducted several experiments on multi-spectral images to understand the impact of unlabeled samples on the classification performance. Our study shows that though in general classification accuracy improves with the addition of unlabeled training samples, it is not guaranteed to get consistently higher accuracies unless sufficient care is exercised when designing a semi-supervised classifier.

Keywords: MAP, MLE, EM, GMM, semi-supervised learning

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. There is a great demand for accurate land use and land cover classification derived from remotely sensed data in various applications. However, increasing spatial and spectral resolution puts several constraints on supervised classification. The increased spectral resolution requires a large amount of accurate training data. Collecting ground truth data for a 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, often supervised learning is carried out with small training samples, which leads to large variance in parameter estimates and thus higher classification error rates. However, a large number of training samples without labels are always available for classification of remote sensing images.

Recently, semi-supervised learning techniques that utilize large unlabeled training samples in conjunction with small labeled training data are becoming popular in machine learning and data mining [8, 6, 9]. This popularity can be attributed to the fact that several of these studies have reported improved classification and prediction accuracies, and that the unlabeled training samples comes almost for free. This is also true in case of remote sensing classification, as collecting samples is free, however assigning labels to them is not. However, it was not clear whether semi-supervised learning improves classification accuracies or not. In this work we developed a method that utilizes unlabeled samples in supervised learning framework and did extensive experimental studies to understand the usefulness of unlabeled training samples in remote sensing imagery classification.

Related Work and Our Contributions: Supervised methods are extensively used in remote sensing imagery classification [12, 7]. Several approaches can be also be found in the literature that specifically deal with small sample size problems in supervised learning [4, 5, 11, 10, 15, 14]. 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 that 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, 6, 9, 2]. The semi-supervised learning tech-

Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)
1082-3409/05 $20.00 © 2005 IEEE

niques have not been well explored in the remote sensing and GIS domains. Only notable study is reported in [13] for hyperspectral data analysis. The common thread between many of these methods is the Expectation Maximization (EM) algorithm. The EM algorithm, first proposed in [3], has become one of the most popular methods for maximum likelihood (ML) based parameter estimation from incomplete data. Key feature of the EM algorithm is that it estimates parameters in the absence of feature values in the input data (also known as incomplete data). Many of the semi-supervised learning methods pose class labels as the missing data and use EM algorithm to improve initial (either guessed or estimated from small labeled samples) parameter estimates.

In text data mining, often it is assumed that the features (words) are independent [9], which leads to simpler statistical models. Often features (spectral bands) in remote sensing imagery are highly correlated, which leads to the assumption of multivariate normal distributions with general covariance matrices. This assumption increases the number of parameters to be estimated. In this paper we provided a new semi-supervised learning method based on expectation maximization (EM) algorithm. As features are highly correlated, we use a Gaussian mixture model (GMM) for describing the training samples and use explicit formulas for estimating all model parameters. In addition, we borrowed the weighting scheme proposed in [9] to weight labeled and unlabeled samples differently in the learning process.

Another objective of this study is to understand the effectiveness of semi-supervised learning with unlabeled samples for multi-spectral remote sensing image classification. Towards this, we have conducted several experiments to evaluate the usefulness of this method in thematic information extraction from multispectral remote sensing imagery.

Paper organization: The rest of this paper is organized as follows. In Section 2, we provide a basic statistical framework for Bayesian classification and maximum likelihood based parameter estimation. In Section 3, we present our semi-supervised learning scheme. Experimental results are given in Section 4, followed by conclusions and future directions in Section 5.

2 Statistical classification framework

In the classification of a remote sensing image, our objective is to assign a class label (y) to each pixel (x – a feature vector) based on a certain decision criterion. Maximum likelihood classification (ML) and maximum a posteriori (MAP) are two of the most widely used statistical classification schemes in remote sensing, which are based on the Bayesian decision theory.

Bayesian Classification: In the Bayesian approach, the objective is to find the most probable set of class labels given the data (feature) vector and *a priori* or prior probabilities for each class. Formally, we can state Bayes' formula as: $P(y_i|x) = \frac{p(x|y_i)P(y_i)}{p(x)}$. Bayes' formula allows us to compute the posterior probability ($P(y_i|x)$) provided that we know the class conditional probability density ($p(x|y_i)$) and the *a priori* probability distribution ($P(y_i)$). The term $p(x)$ is often called the evidence factor, that is, the probability of finding a feature vector x from any of M. The evidence $p(x)$ acts as a scale factor that guarantees that the posterior probabilities sum to one; it has no consequence on the decision rule and is thus often omitted from the decision rule. For a two class (y_1, y_2) problem, the Bayes' decision rule is given by: decide y_1 if $P(y_1|x) > P(y_2|x)$; otherwise decide y_2.

Parameter estimation: We can compute the class conditional densities, $p(x|y_i)$, by assuming suitable parametric model, such as, multivariate normal or Gaussian density. This assumption reduces the difficult problem of estimating an unknown density function $p(x|y_i)$ into a simpler parameter (Θ) estimation problem. Here we use a well-known parameter estimation technique, maximum likelihood estimation (MLE), to obtain the parameter vector Θ from the training samples. First, let us assume that the given training dataset, D, contains n random samples, $x_1, \ldots, x_N$, drawn independently from the pdf $p(x|\theta)$. Then $p(D|\theta)$ is given by, $p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$ The $p(D|\theta)$ in the above equation is also known as the *likelihood* function of θ with respect to the data D (set of training samples for a given class). The *likelihood* function is often represented by the symbol $l(\theta)$ or by $l(\theta|D)$. The MLE of θ is the parameter ($\hat{\theta}$) that maximizes the *likelihood* function $p(D|\theta)$, and is given by, $\hat{\theta} = arg\max_{\theta} \prod_{k=1}^{n} p(x_k|\theta)$. Often it is mathematically simpler to deal with the *log-likelihood* function, $l(\theta) = \ln p(D|\theta)$. Since the $\ln$ function is monotonically increasing, the parameter θ that maximizes the *likelihood* function also maximizes the *log-likelihood* function.

3 Our Approach (Semi-supervised Learning).

In this section, we reformulate the likelihood estimation in the context of finite mixture models and describe a parameter estimation technique that is based on expectation maximization algorithm. This framework also utilizes unlabeled training samples. 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 $D_1, \ldots, D_M$ in some proportions $\alpha_1, \ldots, \alpha_M$, respectively, where $\sum_{i=1}^{M} \alpha_i = 1$ and $\alpha_i \geq 0 (i = 1, \ldots, M)$. Now we can model the data $D = \{x_i\}_{i=1}^{n}$ as being generated independently from the following mixture density: $p(x_i|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j)$

Here $p_j(x_i|\theta_j)$ is the pdf corresponding to the

mixture j and parameterized by θ_j, and $\Theta = (\alpha_1, \ldots, \theta_M, \theta_1, \ldots, \theta_M)$ 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: $L(\Theta) = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j) \right]$. In general, this equation 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* equation can then be simplified as: $L(\Theta) = \ln(P(X,Y)|\Theta)) = \sum_{i=1}^{n} \ln(P(x_i|y_i)P(y)) = \sum_{i=1}^{n} \ln(\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i}))$.

In many supervised learning situations, the class labels (y_i)'s are not readily available. However, assuming that the initial parameters Θ^k can be guessed (as in clustering), or can be estimated (as in semi-supervised learning), we can easily compute the parameter vector Θ using the expectation maximization algorithm. 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. 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 $E[.]$, which is denoted by p_{ij}, is nothing but the probability that Gaussian mixture j generated the data point i, and is given by:

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

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:

$$\hat{\alpha}_j^k = \frac{1}{n}\sum_{i=1}^{n} p_{ij} \quad \text{and} \quad \hat{\mu}_j^k = \frac{\sum_{i=1}^{n} x_i p_{ij}}{\sum_{i=1}^{n} p_{ij}} \qquad (2)$$

$$\hat{\Sigma}_j^k = \frac{\sum_{i=1}^{n} p_{ij}(x_i - \hat{\mu}_j^k)(x_i - \hat{\mu}_j^k)^t}{\sum_{i=1}^{n} p_{ij}} \qquad (3)$$

More detailed derivation of these equations can be found in [1]. Standard semi-supervised algorithms obtain initial estimates of the parameters using the labeled samples D_l, and then uses EM algorithm (equations 2- 3) and unlabeled samples D_{ul} to refine the initial estimates. However, we derived slightly different equations which allows one to use D_l throughout the EM iterations. The new formulation also allows to weight D_l and D_{ul} differently. First, we note that for any two constants, a and b, two correlated random variables can be combined, such that, $E(aX + bY) = a\mu_X + b\mu_Y$. By treating X and Y random variables as D_l and D_{ul}, and constants a and b as different weights, one can emphasize (or deemphasize) the importance of unlabeled samples in the semi-supervised learning using our formulation.

4 Experimental Results

We used a spring Landsat 7 scene, taken on May 31, 2000 over the Cloquet town located in Carlton County of Minnesota state. We designed four different experiments to understand the size and quality of initial labeled samples on the performance of semi-supervised learning, and the impact of unlabeled samples generated from random sampling and informed sampling methods. For all these experiments the test dataset was fixed and consisted of 85 plots. Initial labeled and unlabeled samples were varied as explained in each experiment. From each plot, we extracted exactly 9 feature vectors by centering a 3×3 window on the plot center.

We have two groups of experiments (1,2 and 3,4). Each of these experiments are described below in more detail. In the first group of experiments (1,2) we have about 100 labeled samples which are divided into various subsets of different sizes and a fixed set of 85 unlabeled samples. In all the experiments (1 to 4), we used a fixed test dataset consisting of 85 labeled samples. For discussion purposes we summarized key results as graphs for easy understanding.

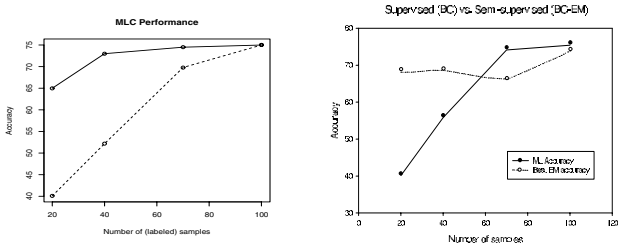
Experiment 1. For this experiment, we generated 5 disjoint labeled training sets, each set consisting of 20 plots at 2 plots per class. We have a fixed unlabeled training dataset consisting of 85 plots.

Experiment 2. For this experiment, we combined 2 sets of labeled samples at a time from the previous experiment to form ${}^5C_2 = 10$ labeled datasets, each consisting of $20 + 20 = 40$ plots. In a similar fashion, we combined 3 different datasets at a time from the above 10 datasets to obtain 3 datasets, each consisting of 70 labeled sample plots (after eliminating duplicate plots).

Experiment 3. The objective of this experiment was to understand the quality and quantity of unlabeled training samples and their impact on overall performance of semi-supervised learning. For this experiment we devised two sampling schemes, simple random sampling, and informed sampling. For the simple random sampling, we generated 10 datasets, each consisting of multiples of 100 sample

plots. No labels were available for these plots. For labeled sample plots we chose two datasets from the first experiment (best [B20] and worst [W20] in terms of MLC accuracies).

Experiment 4. We used informed sampling to generate about 300 unlabeled sample plots. By informed sampling we mean generating random samples in a constrained way using additional information (e.g., existing land-use or land-cover maps, ecological zone maps, population density, clustered or classified image using only labeled samples). These plots were then randomly divided into 4 partitions. The first subset consists of 5 independent training sets, each consisting of 30 plots; second subset consists of 5 training datasets, each consisting of 60 unlabeled plots. Third experiment consists of 3 training datasets, each consisting of 110 unlabeled plots and finally the fourth experiment consists of 2 training datasets each consisting of 170 unlabeled plots. For labeled training we used the same two datasets that were used in experiment 3. For each of these labeled training datasets, semi-supervised learning was carried out against each of the unlabeled training datasets from the above 4 partitions.



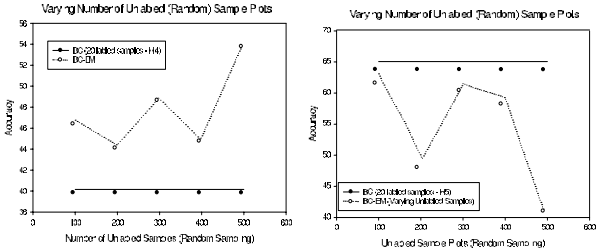
(a) MLC (b) SSL

Figure 1. Classification Performance as the number of (labeled) training samples increases (a) MLC, (b) Semi-supervised.

4.1 Discussion

From the first experiment it is clear that maximum likelihood estimates are highly dependent on both the quantity and the quality of labeled training samples. The plot in Figure 1(a) shows that as the number of training (labeled) samples increases, the conventional maximum likelihood estimates gets better and hence the classification performance of the Maximum likelihood classifier (BC) also improves. It is also interesting to note that the difference between best and worst accuracies gets reduced as the number of samples increase. This is because the noise averages out as the number of samples increases.

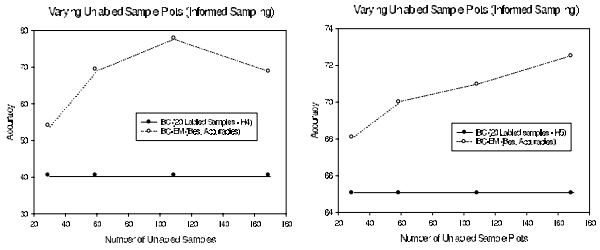
The second experiment shows that as the number of labeled samples increases the usefulness of unlabeled samples diminishes (see Figure 1(b)). Thus the main benefit of semi-supervised learning occurs when there is only a small number of labeled samples available for training.



(a) Against W20 (b) Against B20

Figure 2. Performance of semi-supervised classification as the number of unlabeled samples increases (random sampling).

In next two experiments we explore the impact of the number unlabeled training samples and how they are generated. Figure 2(a) and (b) provides the comparison of randomly generated unlabeled training plots against best and worst cases (labeled training data) taken from the experiment 1. On the other hand Figure 3(a) and (b) shows the results against unlabeled training plots generated by informed sampling. From these two experiments it is clear that accuracy increases as the number of unlabeled training samples increase, however pure random samples might degrade performance quite considerably. The main problem we noticed is that random sampling did not generate enough samples for small (geographic area) classes, as a result the corresponding covariance matrices are becoming singular or close to singular, and the mixing coefficients α_i are close to zero. On the other hand equal (or in proportion to class area) number of samples were generated for each class. It can be seen from the figure that the semi-supervised learning using informed sampling generated unlabeled training plots performed consistently well.



(a) Against W20 (b) Against B20

Figure 3. Performance of semi-supervised classification as the number of unlabeled samples increases (informed sampling).

5 Conclusion and Future Directions

In this study first we presented a semi-supervised learning algorithm for classification of multi-spectral remote sensing imagery. The semi-supervised method presented here uses the classical EM algorithm to augment unlabeled samples to improve initial estimates generated using a small set of training samples. Except for pure randomly generated unlabeled training samples, the semi-supervised learning showed an improved performance in many of the experiments. The overall accuracies varied between -8.67% and $+27.07\%$, and on an average the semi-supervised learning method showed an improvement of 8% in overall accuracy. Given the fact that this is a multi-class (10 classes) classification problem, the accuracies are higher than one would expect from coarse multi-spectral resolution images. This method is very useful in remote sensing data mining, as collection of sufficient training samples for supervised learning is often difficult and costly. However, we also note that getting consistently higher accuracies are not guaranteed with semi-supervised learning method described in this paper. Sufficient care should be taken when selecting the labeled samples as the EM algorithm for Gaussian mixtures is not guaranteed to converge to global optimum. Similarly, appropriate sampling scheme should be employed, such as informed sampling described in this paper, when selecting unlabeled training samples.

Further classification improvements can be expected by incorporating additional GIS layers like population density, upland and lowland maps, digital elevation models, and soil maps. However, these additional layers doesn't follow multivariate normal distribution. We are working on developing a mixture model that admits both continuous random variables and discrete random variables. Also in problem formulation we assumed that the samples follow an i.i.d. distribution, however this assumption is not valid in images as pixels are often spatially auto-correlated. We are working on extending this semi-supervised algorithm to model spatial context in the learning process.

6 Acknowledgments

This research has been supported in part by the Army High Performance Computing Research Center under the auspices of Department of the Army, Army Research Laboratory Cooperative agreement number DAAD19-01-2-0014, and by the cooperative agreement with NASA (NCC 5316) and by the University of Minnesota Agriculture Experiment Station project MIN-42-044. We are particularly grateful to our collaborator Prof. Joydeep Ghosh for useful conversations and critical inputs. We would like to thank Kim Koffolt for improving the readability of this report.

References

[1] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report, University of Berkeley, ICSI-TR-97-021, 1997., 1997.

[2] F. Cozman, I. Cohen, and M. Cirelo. Semi-supervised learning of mixture models. In *Twentieth International Conference on Machine Learning (ICML)*, 2003.

[3] A. 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.

[4] 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*, pages 1–7., 2000.

[5] K. Fukunaga and R. R. Hayes. Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):252–264, 1989.

[6] S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proc. 17th International Conf. on Machine Learning*, pages 327–334. Morgan Kaufmann, San Francisco, CA, 2000.

[7] J. R. Jensen. *Introductory Digital Image Processing, A Remote Sensing Perspective*. Prentice Hall, Upper Saddle River, NJ-07458, 1996.

[8] T. Mitchell. The role of unlabeled data in supervised learning. In *Proceedings of the Sixth International Colloquium on Cognitive Science, San Sebastian, Spain.*, 1999.

[9] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[10] S. Raudys. On dimensionality, sample size, and classification error of nonparametric linear classification algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(6):667–671, 1997.

[11] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):252–264, 1991.

[12] J. A. Richards and X. Jia. *Remote Sensing Digital Image Analysis*. Springer, New York, 1999.

[13] 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*, 32(5), 1994.

[14] M. Skurichina and R. Duin. Stabilizing classifiers for very small sample sizes. In *Proc. 10th Int. Conference on Pattern Recognition, IEEE Computer Society Press*, pages 891–896, 1996.

[15] S. Tadjudin and D. A. Landgrebe. Covariance estimation with limited training samples. *IEEE Trans. Geosciences and Remote Sensing.*, 37(4):2113–2118, 1999.