Adopting Semi-supervised Learning Algorithms for Mining Remote Sensing Imagery: Summary of Results and Open Research Problems

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

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 multispectral 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. We also extended this semi-supervised framework to model spatial context through Markov Random Fields and initial experiments shows an improved accuracy over MLC, Semi-supervised, and MRF classifiers. Though this study shows that semi-supervised learning schemes can be adopted for remote sensing data mining, there are some open research issues that needs to be solved before these methods can be applied in production environments.

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. On the other hand increased spatial resolution mandates modeling neighborhood (context) relationships in classification. 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 [12, 8, 13]. 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 almost 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. As the spatial context is also important for improving classification accuracy and reduce 'salt and pepper' noise, we extended this semi-supervised learning framework via Markov Random Fields (MRF). This paper summarizes the initial results and discusses some open research problems.

Related Work and Our Contributions: Supervised methods are extensively used in remote sensing imagery classification [18, 10]. Several approaches can be also be found in the literature that specifically deal with small sample size problems in supervised learning [6, 7, 17, 16, 23, 21]. These methods are aimed at designing appropriate clas-

sifiers, 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 [12, 8, 13, 4]. The semi-supervised learning techniques have not been well explored in the remote sensing and GIS domains. Only notable study is reported in [19] for hyperspectral data analysis. The common thread between many of these methods is the Expectation Maximization (EM) [5] algorithm. 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 [13], 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.

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 multi-spectral remote sensing imagery. Finally, we extended this semi-supervised learning scheme via MRF to model spatial context.

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 ($P(y_i)$). Two popular Bayesian classifier, MLC and MAP are given below. We use these two classifier as basis for our semi-supervised learning.

$$g_i(x) = -\ln|\Sigma_i| - (x-\mu_i)^t|\Sigma_i|^{-1}(x-\mu_i) \quad (1)$$

$$g_i(x) = \ln P(y_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{-1}{2}(x-\mu_i)^t|\Sigma_i|^{-1}(x-\mu_i) \quad (2)$$

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 Semi-supervised Learning

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:

$$p_{ij} = \frac{\left|\hat{\Sigma}_j\right|^{-1/2} e^{\left\{-\frac{1}{2}(x_i-\hat{\mu}_j)^t \hat{\Sigma}_j^{-1}(x_i-\hat{\mu}_j)\right\}}}{\sum_{l=1}^{M}\left|\hat{\Sigma}_l\right|^{-1/2} e^{\left\{-\frac{1}{2}(x_i-\hat{\mu}_l)^t \hat{\Sigma}_l^{-1}(x_i-\hat{\mu}_l)\right\}}} \quad (3)$$

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:

$$\hat{\alpha}_j^k = \frac{1}{n}\sum_{i=1}^{n} p_{ij} \quad \text{and} \quad \hat{\mu}_j^k = \frac{\sum_{i=1}^{n} x_i p_{ij}}{\sum_{i=1}^{n} p_{ij}} \quad (4)$$

$$\hat{\Sigma}_j^k = \frac{\sum_{i=1}^{n} p_{ij}(x_i-\hat{\mu}_j^k)(x_i-\hat{\mu}_j^k)^t}{\sum_{i=1}^{n} p_{ij}} \quad (5)$$

More detailed derivation of these equations can be found in [3]. Standard semi-supervised algorithms obtain initial estimates of the parameters using the labeled samples D_l, and then uses EM algorithm (equations 4- 5) and unlabeled samples D_{ul} to refine the initial estimates. However, we derived slightly different update equations which allows one to use D_l (as they are the most representative training samples) throughout the EM iterations. The new formulation also allows us 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. The new equations are given by:

$$\hat{\alpha}_j^k = \frac{(\lambda_l m_j + \sum_{i=1}^{n}\lambda_{ul} p_{ij})}{(\lambda_l m + \lambda_{ul} n)} \quad (6)$$

$$\hat{\mu}_j^k = \frac{(\sum_{i=1}^{m_j}\lambda_l y_{ij} + \sum_{i=1}^{n}\lambda_{ul} x_i p_{ij})}{(\lambda_l m_j + \sum_{i=1}^{n}\lambda_{ul} p_{ij})} \quad (7)$$

$$\hat{\Sigma}_j^k = \frac{\left\{\begin{array}{l}\sum_{i=1}^{m_j}\lambda_l(y_{ij}-\hat{\mu}_j^k)(y_{ij}-\hat{\mu}_j^k)^t + \\ \sum_{i=1}^{n} p_{ij}\lambda_{ul}(x_i-\hat{\mu}_j^k)(x_i-\hat{\mu}_j^k)^t\end{array}\right\}}{(\lambda_l m_i + \sum_{i=1}^{n}\lambda_{ul} p_{ij})} \quad (8)$$

4 Contextual Semi-supervised Learning

There are two major approaches for modeling spatial dependencies (context, neighborhood relationships, spatial autocorrelation) in prediction/classification problems, namely, spatial autoregressive models (SAR) and Markov random fields (MRF). These two models were compared in Shekhar et al. [20]. Knowledge discovery techniques which ignore spatial autocorrelation typically perform badly on the spatial datasets. Over the last decade, several researchers [22], [11], [24] have exploited spatial context in classification using Markov Random Fields to obtain higher accuracies over their counterparts (i.e., non-contextual classifiers). MRFs provide a uniform framework for integrating spatial context and deriving the probability distribution of interacting objects.In this paper we extended the semi-supervised learning algorithm (Section 3) to model spatial context via the MAP-MRF model. MRF exploits spatial context through the prior probability $p(y_i)$ term in the Bayesian formulation (Section 2). For a Markov Random Field y, the conditional distribution of a point in the field given all other points is only dependent on its neighbors; given as

$$p(y(i,j)|y(k,l);\ k,l \neq i,j\ =\ p(y(i,j)|y(k,l);\ k,l \in s). \quad (9)$$

where s is the local neighborhood of pixel at (i,j). Now the problem is how to incorporate this MRF locality property into the MAP solution given in eq. 2 Gibs Random Fields (GRF) provide an easy way of incorporating this neighborhood information. GRFs are defined in terms of a joint distribution of random variables, which is easier to compute, as opposed to the conditional distribution given by MRFs. Gibs distribution for a given clique is defined as:

$$p(y)\ =\ \frac{1}{z} e^{-\frac{1}{T}\Sigma_C V_c(y)} \quad (10)$$

where $V_c(y)$ is the potential associated with clique c, and C is the set of all cliques. According to the Hammersley-Clifford theorem [1], there is a one-to-one correspondence between MRFs and GRFs. Therefore, if $p(y)$ is formulated as a Gibbs distribution, y should have the properties of a Markov random field. Since, MRF models spatial context in the *a priori* term, we optimize a *penalized log-likelihood* [9] instead of the *log-likelihood* function defined in Section 3. The *penalized log-likelihood* can be written as

$$ln(P(X,Y|\Theta)) = -\sum_{C} V_C(y,\beta) - lnC(\beta) + \sum_i \sum_j Y_{ij} \ln p_j(x_i|\Theta_i) \quad (11)$$

Then the E-step for a given Θ^k, reduces to computing

$$Q(\Theta, \Theta^k) = \sum_i \sum_j E(Y_{ij}|x, \theta^k) ln p_j(x_i|\theta_i) \quad (12)$$
$$- \sum E(Vc(Y, \beta)|x, \theta^k) - lnC(\beta)$$

However, exact computation of the quantities $E(Vc(Y, \beta)|x, \theta^k)$ and $E(Y_{ij}|x, \theta^k)$ in the eq. 12 are impossible [14]. Also the maximization of eq. 12 with respect to β is also very difficult, because of computing $z = C(\beta)$ is intractable except for very simple neighborhood models. Several approximate solutions for this problem in un-supervised learning can be found in [14, 15]. We extended the approximate solution provided in [14] for semi-supervised learning and showed its usefulness in improving land cover and land use predictions from remote sensing imagery. The E-step is divided into two parts: first, we compute complete data *log-likelihood* for all data points, second, for the given neighborhood, we iteratively optimize contextual energy using iterative conditional modes (ICM) [2] algorithm. Since the estimation of β the is difficult [14], we assume that it is given *a priori*, and proceed with M-step as described in the semi-supervised learning algorithm.

5 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.

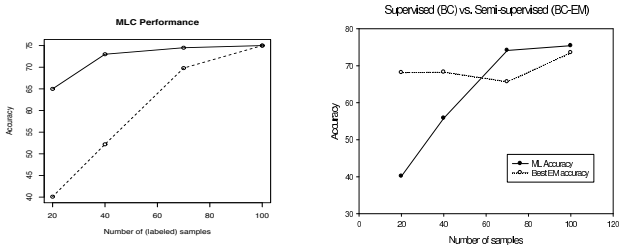
We have two groups of experiments (1,2 and 3,4). Each of these experiments are described below in more detail. In the first group of experiments (1,2) we have about 100 labeled samples which are divided into various subsets of different sizes and a fixed set of 85 unlabeled samples. In all the experiments (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).

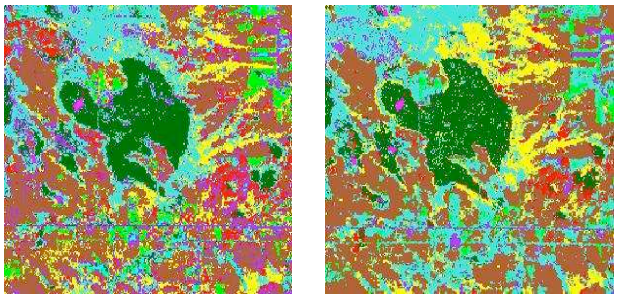
Experiment 4. We used informed sampling to generate about 300 unlabeled sample plots. By informed sampling we mean generating random samples in a constrained way using additional information (e.g., existing land-use or land-cover maps, ecological zone maps, population density, clustered or classified 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.



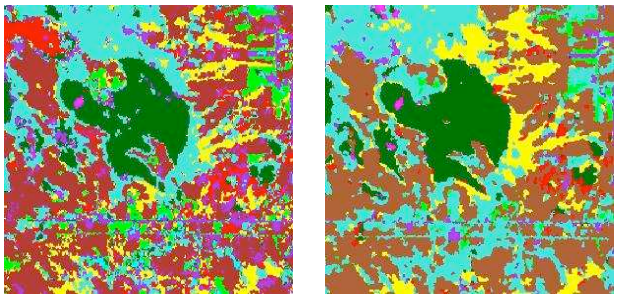
(a) MLC (b) SSL

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

Experiment 5. This experiment consists of applying all four



(a) MAP (b) Semi-supervised



(c) MRF (d) Contextual Semi-supervised

Figure 2. Small portion from the classified NW corner of the Carleton image. (a) Bayesian (MAP), (b) Semi-supervised (EM-MAP), (c) MRF (MAP-MRF) and (d) Contextual Semi-supervised (EM-MAP-MRF)

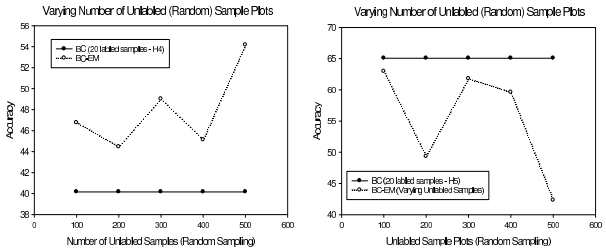
classifiers, namely, MAP, Semi-supervised, MAP-MRF, and Contextual Semi-supervised. The results were summarized in the Figure 2.

5.1 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.

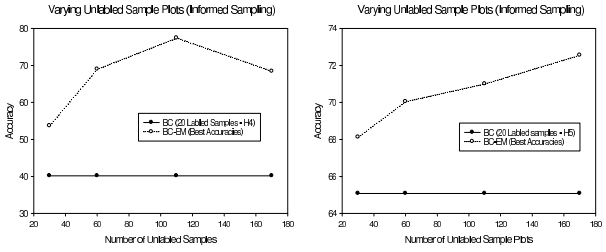
In next two experiments we explore the impact of the



(a) Against W20 (b) Against B20

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

number unlabeled training samples and how they are generated. Figure 3(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 4(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 4. Performance of semi-supervised classification as the number of unlabeled samples increases (informed sampling).

6 Conclusion and Open Research Problems

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.

From the Figure 2(b), it can be seen that though semi-supervised learning is more accurate than the base MAP classifier, the classified image contains lot of 'salt and pepper' noise. It should also be noted from the Figures 2(c) and (d) that modeling context is classification not only improves the overall accuracy but also eliminates the 'salt and pepper' noise. The output of contextual semi-supervised classification is more desirable from several other GIS applications point of view.

Further research is needed for incorporating additional GIS layers like population density, upland and lowland maps, digital elevation models, and soil maps into the semi-supervised learning. Right now there are no suitable statistical model available that can handle these heterogeneous attributes. We are working on developing a mixture model that admits both continuous random variables and discrete random variables. We also identified two issues with contextual semi-supervised learning, namely, performance and convergence. In all our experiments, the contextual semi-supervised learning converged, however, formal theoretical proof of convergence is need. A close look at the contextual semi-supervised algorithm, reveals that the contextual energy is optimized at each iteration of the EM algorithm, which is clearly not desirable from the computational complexity point of view. We need smarter algorithms to speedup the convergence and as well reduce the need to optimize contextual energy at each iteration. Further research is also needed to develop other approximate solutions, such as, linear programming and graph min-cut algorithms.

7 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. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society*, 36:192–236, 1974.

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society*, 48(3):259–302, 1986.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] P. J. Green. On use of the em algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B*, 52(3):443–452, 1990.

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

[11] Y. Jhung and P. H. Swain. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(1):67–75, 1996.

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

[13] 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.

[14] W. Qian and D. Titterington. Estimation of parameters in hidden markov models. *Philosophical Transactions of the Royal Statistical Society, Series A*, 337:407–428, 1991.

[15] W. Qian and D. Titterington. Stochastic relaxations and em algorithms for markov random fields. *Journal of Statistical Computation and Simulation*, 41, 1991.

[16] 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.

[17] 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.

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

[19] 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.

[20] S. Shekhar, P. Schrater, R. Vatsavai, W. Wu, and S. Chawla. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transaction on Multimedia*, 4(2):174–188, 2002.

[21] 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.

[22] A. H. Solberg, T. Taxt, and A. K. Jain. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing*, 34(1):100–113, 1996.

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

[24] C. E. Warrender and M. F. Augusteijn. Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, 20(10):1987–2002, 1999.