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

Ranga Raju Vatsavai¹,³, Shashi Shekhar¹, and Thomas E. Burk².

¹Spatial Database Research Group, Department of Computer Science.

²Remote Sensing and Geospatial Analysis Lab, Department of Forest Resources, University of Minnesota.

³IBM-Research, India Research Laboratory, Indian Institute of Technology-Delhi, India.
Outline

- Introduction
- Semi-supervised Classification
- Results
- Conclusions and Open Research Problems
Introduction

- Remote Sensing
  - Spectral Resolution – multi-spectral to hyper-spectral
  - Spatial Resolution – low (60m) to high (1m)
Introduction

- Classification*
  - Supervised – MLC, MAP, DT, NN, ...
  - Unsupervised – Clustering (kNN, kMeans, GMM...)

Supervised Classification Process

- Example Application - Thematic Mapping
  - 7 Bands, 7000Px7000L, 30m Pixel size, 16 day revisit, 170x183 km.

Sample plots
ID, Lat, Lon

Training Data
ID, \( b_1, \ldots, b_n \), Label
Supervised Classification Process

\[ x \in X \]

\[ \begin{align*}
  \text{Sampling} & \quad \text{Labeled } D^{l} \\
\end{align*} \]

\[ \begin{align*}
  \text{Training Data} & \quad \text{Classifier Model} \\
\end{align*} \]

\[ p(\omega_i | x) = \frac{p(x | \omega_i) \cdot p(\omega_i)}{p(x)} \]

\[ N(\mu, \Sigma) \]
Supervised Classification Process

- **Bayes’ Theorem:**
  \[ p(\omega_i | x) = \frac{p(x | \omega_i) \cdot p(\omega_i)}{p(x)} \]

- Assuming \(D_i = \{x_1, \ldots, x_n\}\), and x’s i.i.d, the likelihood function is simply:
  \[ p(D_1 | \theta_1) = \prod_{k=1}^{n} p(x_k | \theta_1) \]

- The maximum-likelihood estimation of \(\theta_1\) is the value that maximizes \(p(D_1 | \theta_1)\).
  \[ \hat{\theta}_1 = \arg \max_{\theta_1} l(\theta_1) \]
  \[ \hat{\mu}_1 = \frac{1}{n} \sum_{k=1}^{n} x_k \]
  \[ \hat{\Sigma}_1 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu}_1)(x_k - \hat{\mu}_1)' \]

- MLE gives familiar quantities:
  - Asymptotically unbiased, consistent, efficient
  - All these properties are true if \(n \rightarrow \infty\)
  - \(|D| \sim (10 \text{ to } 100) \times \text{(number of dimensions)}\)
Supervised Classification Process

- MLC performance as $|D|$ increases

![MLC Performance Chart]
**Symbols**

$x = \text{Feature Vector}$

$n = \# \text{of features or } \# \text{dimensions}$

$\mu = \text{Mean Vector}$

One $\mu$ per class

$\Sigma = \text{Covariance Matrix}$

One $\Sigma$ per class

Subscripts are context dependent

\[
x = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}_{nx1} = \begin{bmatrix} 63 \\ 53 \\ 35 \\ 121 \\ 76 \\ 31 \end{bmatrix}, \quad \mu_\omega = \begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix}_{nx1} = \begin{bmatrix} 67.33 \\ 60.44 \\ 41.44 \\ 89.77 \\ 79.66 \\ 40.44 \end{bmatrix}
\]

\[
\Sigma_\omega = \begin{bmatrix} v_{1,1} & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & v_{n,n} \end{bmatrix}_{nxn} = \begin{bmatrix} 2.0 & 0.2 & 0.33 & 0.58 & 1.13 & -0.7 \\ 0.41 & 1.7 & 0.6 & 0.2 & -0.3 & -1.2 \\ -1.16 & & & & & \\
2.7 & & & & & 
\end{bmatrix}
\]
Semi-supervised Classification

- Acquiring Labels
  - Costly, Time consuming, Labor intensive
  - Error prone
  - May be impossible at times –
    - Emergency situations (fires, floods, cyclones, …)
    - Accessibility, Privacy

- Samples are readily available, but not labels
- There are several classification problems that have to deal with insufficient learning samples
- Q? Can we still learn with partially labeled training data
Semi-supervised Classification Process

Classification Algorithms

\[
p(\omega_i \mid x) = \frac{p(x \mid \omega_i) \cdot p(\omega_i)}{p(x)}
\]
Assume current estimate of parameter is \( \theta_i \)

See what happens to \( L \) when new \( \theta \) is estimated

\[
L(\theta) - L(\theta_i) = \ln p(x | \theta) - \ln p(x | \theta_i) = \ln \frac{p(x | \theta)}{p(x | \theta_i)}
\]

We would like to choose \( \theta \) to maximize r.h.s

EM*, introduce unobserved vars \( Z \) such that if \( Z \) is known, the \( \theta \) is computed easily

\[
L(\theta) - L(\theta_i) = \ln \sum_z p(x | z, \theta) p(z | \theta) \frac{p(x | \theta)}{p(x | \theta_i)}
\]

The EM algorithm first finds the expected value of the complete-data log-likelihood –

- E-step

\[ Q(\theta; \theta^{(i)}) = E[\log p(D^l, D^u; \theta) | D^l; \theta^{(i)}] \]

- The second step (M-step) of the EM algorithm is to maximize the E[] of 1st step.

- M-step

\[ \theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)}) \]
EM update equations for GMM*

- Assume that D is drawn from a Gaussian Mixture Model, given by

\[ p(x | \theta) = \sum_{i=1}^{M} \alpha_i p_i(x | \theta_i) \]

where \( \theta = (\alpha_1, ..., \alpha_M; \theta_1, ..., \theta_M) \) such that \( \sum_{i=1}^{M} \alpha_i = 1, \ 0 < \alpha_i < 1 \) and \( p_i \) pdf parameterized by \( \theta_i \)

GMM – Update Equations

- **E-Step**

\[
e_{ij} = \frac{\left| \hat{\Sigma}_k \right|^{-1/2} \exp \left\{ -\frac{1}{2} \left(x_i - \hat{\mu}_j^k \right)^T \hat{\Sigma}_k^{-1,k} \left(x_i - \hat{\mu}_j^k \right) \right\}}{\sum_{l=1}^M \left| \hat{\Sigma}_l \right|^{-1/2} \exp \left\{ -\frac{1}{2} \left(x_i - \hat{\mu}_l^k \right)^T \hat{\Sigma}_l^{-1,k} \left(x_i - \hat{\mu}_l^k \right) \right\}}
\]

- **M-Step**

\[
\alpha_j = \frac{\sum_{i=1}^N e_{ij}}{N} , \quad \hat{\mu}_j^{k+1} = \frac{\sum_{i=1}^N e_{ij} x_i}{\sum_{i=1}^N e_{ij}} , \quad \hat{\Sigma}_j^{k+1} = \frac{\sum_{i=1}^N e_{ij} \left(x_i - \hat{\mu}_j^k\right) \left(x_i - \hat{\mu}_j^k\right)^T}{\sum_{i=1}^N e_{ij}} \quad \text{i}^{\text{th}} \text{data vector, } j^{\text{th}} \text{ class}
\]
GMM – New Update Equations

\[ \alpha_j = \frac{\left( \lambda_l m_j + \lambda_u \sum_{i=1}^{n} e_{ij} \right)}{\lambda_l m + \lambda_u n}, \]

\[ \hat{\mu}_j^{k+1} = \frac{\left( \lambda_l \sum_{i=1}^{m_j} y_{ij} + \lambda_u \sum_{i=1}^{n} e_{ij} x_i \right)}{\lambda_l m_j + \lambda_u \sum_{i=1}^{n} e_{ij}}, \]

\[ \hat{\Sigma}_j^{k+1} = \frac{\left( \lambda_l \sum_{i=1}^{m_j} (y_{ij} - \hat{\mu}_j^{k+1} )(y_{ij} - \hat{\mu}_j^{k+1})^T + \lambda_u \sum_{i=1}^{n} e_{ij} (x_i - \hat{\mu}_j^{k+1} )(x_i - \hat{\mu}_j^{k+1})^T \right)}{\lambda_l m_j + \lambda_u \sum_{i=1}^{n} e_{ij}}. \]

\(i^{th}\) data vector, \(j^{th}\) class
Semi-supervised learning Algorithm - Outline

- **Inputs:** Collection \( D \) consisting of \( D^l \) and \( D^u \)
- **Build initial classifier,** estimate \( \hat{\theta} = \arg \max_{\theta} l(\theta) \)
- **Loop while parameter estimate improves**
  - (E-Step): Use current classifier estimate, \( \hat{\theta} \), to estimate component membership of each data vector, i.e., the prob. that each mixture component (class) generated each sample
  - (M-Step): Re-estimate the classifier parameter, given the estimated component membership of each sample
- **Output:** A classifier, with improved estimates, that takes any new data sample and predicts a class label.
Experimental Results

**Experiment 1**
- Learning MLE
- **Learning datasets for first two experiments**
  - $|D_1| = 100$, Labeled
  - $|D_1^i| = 20$

**Experiment 2**
- Learning EM
- **Learning datasets for first two experiments**
  - $|D_1| = \{D_1^i\}_{i=1}^{5}$
  - $|D_1^i| = 20$

**Combine 2 partitions at a time**

**Experiment 3**
- Learning MLE
- **Learning datasets for first two experiments**
  - $|D_1| = \{D_1^i\}_{i=1}^{10}$
  - $|D_1^i| = 40$

**Experiment 4**
- Learning EM
- **Learning datasets for first two experiments**
  - $|D_1| = \{D_1^i\}_{i=1}^{3}$
  - $|D_1^i| = 70$

**Test dataset**
- $|D_{te}| = 85$
- Same for Exp 1 & 2

**Learning datasets for experiments 2 and 3**
Two learning datasets from Experiments 1 are chosen based on best (B20) and worst (W20) accuracies

- Random Sampling
  - $D_{ul} = \{D_{ul}^i\}_{i=1}^{9}$
  - $|D_{ul}^i| = i \times 100$

- Repeat Semi-supervised Learning and Testing as described in experiment 1.

- Informed Sampling
  - Total 300 plots & Partitions = 4
  - See corresponding accuracy table for size of each data set.
Experimental Results

MLC Performance

Accuracy vs. Number of (labeled) samples
Experimental Results

Varying Number of Unlabeled (Random) Samples

Varying Number of Unlabeled (Random) Sample Plots
Experimental Results

MLE (100 labeled)

MLE (20 lab)
Experimental Results

MLE (100 labeled)  

EM (20 lab + 85 ul)
Spatial Semi-supervised

- iid assumptions are not valid
- MAP/MRF
  - \( p(c|x) = \frac{p(x|c)P(c)}{p(x)} \)
  - \( P(c) = P(c(i, j) | c(k, l); \{k, l\} \neq \{i, j\}). \)
  - \( = P(c(i, j) | c(k, l); \{k, l\} \in s). \)
  - \( = \frac{1}{Z} e^{-\frac{U(\omega)}{T}} \)

- Z in not computable
  - Assume
    - 10 classes,
    - 256x256 image
  - Total configurations
    - \( 10^{(256 \times 256)} \)

\[ U_s(C(i, j)) = \sum_{\{k, l\} \in s_{ij}} \beta I(C(i, j), C(k, l)) \]

Where

\[ I(C(i, j), C(k, l)) = \begin{cases} 
1 & \text{if } C(i, j) = C(k, l) \\
0 & \text{if } C(i, j) \neq C(k, l)
\end{cases} \]
Spatial Semi-supervised

BC (60%)

BC-EM (68%)

BC-MRF (65%)

BC-EM-MRF (72%)
Open Research Problems

- Care should be exercised when collecting unlabeled samples
  - No samples from small classes
  - Mixed plots
  - Samples from unknown components

- Extend the model for multi-source data
Open Research Problems

- Extend the model for multi-source data
Open Research Problems

- Discovering sub-components (e.g., Anderson Level 2 and 3 classes)
Open Research Problems

- Efficient Algorithms
- Global Optimization (Stochastic versions)
- Semi-supervised MCC
- Semi-supervised + Active Learning

- Not a research problem, but
  - Can we create bench mark dataset(s) similar to CMU or UCI for Remote Sensing Data Mining.

- Conclusions
  - Results were promising
Acknowledgements

- Prof. Joydeep Ghosh, UT Austin
- Team @ Remote Sensing & Geospatial Analysis Lab
- Team @ Spatial Databases Research Group, C.Sci. Dept. U of Minnesota.