AN EXPERIMENTAL COMPARISON OF GRAPH BASED SEMI-SUPERVISED LEARNING FOR MULTISPECTRAL IMAGE CLASSIFICATION

E.M. Tu [a,b], J. Yang [*a,b], J.X. Fang [a,b], Z.H. Jia[c], N. Kasabov [d]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University;
[b] Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai, 200240, China - (tuen, jieyang, fangchj2002)@sjtu.edu.cn
[c] Xinjiang University, School of Information Science and Engineering ,Urumqi, 830046, China - jzhh@xju.edu.cn
[d] The Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, New Zealand - nkasabov@aut.ac.nz

KEY WORDS: Multispectral Image Classification, Semi-Supervised Learning, Graph Laplacian, Multivariate Taylor Expansion, Woodbury Formula, Large Matrix Inversion

ABSTRACT:

Semi-Supervised Learning recently catches much attention and has demonstrated its superiority to classify abundant unlabelled samples with only a few labelled samples. The goal of this paper is to provide an experimental comparison of the efficiency of graph based semi-supervised learning algorithms in context of multispectral image classification. We compared the classification accuracy and the spatial and temporal complexity of several standard graph based semi-supervised learning algorithms. We also propose an efficient semi-supervised learning algorithm which has linear complexity in both temporal and spatial domain. To achieve this, we first use multivariate Taylor series expansion to the Gaussian kernel function and then use the Woodbury formula to convert a large matrix inversion problem to a small matrix inversion problem. Experimental results show that our proposed algorithm can attain high accuracy when only a few training samples are available.

1. INTRODUCTION

Multispectral image classification is a basic problem arising in change detection(Lu, Mausel et al. 2004; Radke, Andra et al. 2005), land use/land cover investigation(Sobrino and Raissouni 2000; Friedl, McIver et al. 2002), urban planning(Pauleit and Duhme 2000), etc. Traditional methods for image classification only use either labelled samples (supervised learning) or unlabelled samples (unsupervised learning).

Graph based Semi-Supervised Learning (SSL), a learning framework between supervised learning and unsupervised learning, can do classification by simultaneously using both labelled and unlabelled samples. In the last decade, SSL has drawn much attention in pattern recognition and computer vision fields(Chapelle, Sch lkopf et al. 2006; Zhu 2006; Zhu and Goldberg 2009) due to its superiority over traditional supervised learning, like Support Vector Machine (SVM), Artificial Neural Networks (ANN), *K* Nearest Neighbours (KNN) etc., for classifying abundant unlabelled samples with only a few labelled samples available. For a continuously updating survey and some reference book, we recommend (Chapelle, Schölkopf et al. 2006; Zhu 2006; Zhu and Goldberg 2009).

In the remote sensing field, the value of graph based semi-supervised learning has not been fully explored. Semi-supervised support vector machine (S3VM) has been studied a lot for the remote sensing images classification (Bruzzone, Chi et al. 2006; Marconcini, Camps-Valls et al. 2009), but graph-based semi-supervised learning, which has been studied deeply in computer vision and machine learning fields because of its solid mathematical background and excellent performance(Shi and Malik 2000; Zhu, Ghahramani et al. 2003; Zhou, Bousquet et al. 2004; Belkin, Niyogi et al. 2006), has been paid little attention in the remote sensing classification. The study of semi-supervised learning in the remote sensing field is particularly important for reasons below: 1) In remote sensing applications sufficient high-quality labelled samples are often difficult to obtain because they need long time, special devices and too many human activities to retrieve or label. On the other hand, however, huge amount of unlabelled data are relatively easy to collect. 2) Traditional supervised learning needs to be trained with the samples which cover the whole class distribution in order to achieve low generalization error. However, for remote sensing tasks, this is not easy because of the presence of transitional classes, restricted access to sites, uncertainties in classes, etc. (Powell, Matzke et al. 2004; Bradley 2009).

The goal of this paper is to provide an experimental comparison of the efficiency of graph based semi-supervised learning in the context of multispectral image classification. We compared several standard graph based semi-supervised learning algorithms, as well as a new one we have recently developed, in respect to the classification accuracy and the spatial and temporal complexity. The algorithms we studied include Harmonic Function (HF)(Zhu, Ghahramani et al. 2003), Local and Global Consistency (LGC) (Zhou, Bousquet et al. 2004), Anchor Graph Regularization (AGR)(Liu, He et al.) and Nystrom Approximation Method (NAM)(Camps-Valls, Bandos Marsheva et al. 2007) as well as our Taylor Series Expansion algorithm(TSE). In many experiments our new algorithm keeps

* Corresponding Author: jieyang@sjtu.edu.cn

comparable classification accuracy, while saves considerable space and time.

2. GRAPH BASED SEMI-SUPERVISED LEARNING

2.1 General Notions of Graph

We first introduce some notions. Given a sample set $V = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k, \mathbf{x}_{k+1}, ..., \mathbf{x}_n \mid \mathbf{x}_i \in \mathbb{R}^d\}$ coming from c classes, and a class label set $\mathcal{L} = \{1, 2, ..., c\}$, we assume that the first k samples in V have their class labels $\{y_1, y_2, ..., y_k \mid y_i \in \mathcal{L}\}$ and the remaining $n-k$ samples' labels are unknown. Define $G = (V, E)$ as a graph with vertex set V and edge set $E = \{e(\mathbf{x}_i, \mathbf{x}_j) \mid 1 \le i, j \le n\}$. Since there is a one-to-one correspondence between graph vertices and samples, we will use both "vertex $\mathbf{x}_i$" and "sample $\mathbf{x}_i$" indifferently, i.e. vertex $\mathbf{x}_i$ and sample $\mathbf{x}_i$ mean the same thing. If vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ have a connection, then $e(\mathbf{x}_i, \mathbf{x}_j) = w_{ij}$; otherwise $e_{ij} = 0$. The weight w_{ij} is a nonnegative real number which measures the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$. The matrix $\mathbf{W} = \{w_{ij} \mid 1 \le i, j \le n\}$ is called weight matrix or adjacency matrix of the graph. The volume of vertex i is defined as $vol(i) = \sum_j w_{ij}$. The Laplacian of graph $G = (V, E)$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with its diagonal elements $D_{ii} = vol(i)$. The normalized Laplacian is defined as $\mathbf{L}_{norm} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{S}$, where $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ is the normalized weight matrix. It can be proved that both $\mathbf{L}$ and $\mathbf{L}_{norm}$ are positive semi-definitive matrices(Chung 1997).

2.2 Graph Based Semi-Supervised Learning

There are two basic assumptions for semi-supervised learning: 1) samples that are close to each other tend to be in the same class; 2) samples that are distributed on same manifold tend to be in the same class. Different formulations of these two assumptions yield several popular algorithms, including the Harmonic Function(Zhu, Ghahramani et al. 2003), Local and Global Consistency(Zhou, Bousquet et al. 2004) and Manifold Regularization(Belkin, Niyogi et al. 2006).

Since it is closely related our algorithm, here we focus on the Local and Global Consistency learning method. Let f be the decision function defined on graph $G = (V, E)$. Then these assumptions mean: *a*) f should be sufficiently smooth on graph G, such that it will not change abruptly on nearby vertices coming from same class; *b*) f should have consistency property, such that it will not conflict with the prior information contained in the labelled samples. Combining these two constraints together, it will produce to semi-supervised learning algorithms in the framework of graph regularization.

Define a $n \times c$ initial labelling matrix $\mathbf{Y}$ (n is the total sample number and c is the number of classes from which these samples are retrieved) and let $Y_{ij} = 1$ if sample $\mathbf{x}_i$ comes from class j; $Y_{ij} = 0$ otherwise. Let $\mathbf{F}$ be a $n \times c$ matrix which we will learn on the graph. Then the Local and Global Consistency learning method defines an objective function as

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2}\left(\sum_{i,j=1}^{n} w_{ij} \left\| \frac{1}{\sqrt{vol(i)}}\mathbf{F}_i - \frac{1}{\sqrt{vol(j)}}\mathbf{F}_j \right\|^2 + \mu \sum_{i=1}^{n} \left\| \mathbf{F}_i - \mathbf{Y}_i \right\|^2 \right) \quad (1)$$

where $\mathbf{F}_i$ and $\mathbf{Y}_i$ are the row vectors of matrices $\mathbf{F}$ and $\mathbf{Y}$ respectively, and μ is a regularization parameter. The first term is smoothness constraint and the second term is consistency constraint. By setting the differentiation of $\mathcal{Q}(\mathbf{F})$ with respect to $\mathbf{F}$ to zero, it will produce

$$\mathbf{F}^* = \beta(\mathbf{I} - \gamma\mathbf{S})^{-1}\mathbf{Y} \quad (2)$$

where $\beta = \mu/(1+\mu), \gamma = 1/(1+\mu)$. Then the class label of sample $\mathbf{x}_i$ is determined by $y_i = \arg\max_{j=1,...,c} F_{ij}^*$. It is not difficult to show that the entry F_{ij}^* has the property: $0 \le F_{ij}^* \le 1$. From this point, the entry F_{ij}^* can be treated as the likelihood of sample $\mathbf{x}_i$ coming from class j.

If there are n samples (including the labelled and unlabelled) coming from c classes, then to obtain the optimal solution using equation(2), one needs to invert a $n \times n$ matrix. This operation has both time and space polynomial complexity. To be worse, in the remote sensing filed, the total sample number n can be easily up to millions or more. For example for a 400×400 TM imagery, the dimension of the adjacency matrix or Laplacian matrix of the graph is 160000×160000 (The memory space consumed by a double precision full matrix is about 46GB. Although the space can be reduced by adopting sparse matrix, the memory is still considerably large when samples are densely distributed or gathered into compact clusters.). It is very intractable to invert such a matrix directly on a PC. Besides, large matrix inversion tends to have large perturbation due to noise affection. Therefore an efficient algorithm has to be developed.

In the next section, we give an algorithm which only needs to invert a small matrix, and thus has linear complexity in space and time.

3. EFFICIENT GRAPH BASED SEMI-SUPERVISED LEARNING

Many graph based SSL algorithms suffer from their polynomial complexity in both time and space, including HF, LGC and MR. Several speed-up algorithms have been reported in literature recently, including Nystroem approximation method(NAM) (Camps-Valls, Bandos Marsheva et al. 2007), Anchor Graph Regularization (AGR)(Liu, He et al.) and the eigenfunction approach(Fergus, Weiss et al. 2009). In (Camps-Valls, Bandos Marsheva et al. 2007) a graph-based semi-supervised learning algorithm adopting Nystroem approximation method is introduced. But from our experience, the Nystroem method tends to be unstable when the chosen rows/columns are highly

correlated. Besides, it also needs to make a trade-off between the performance and the number of rows/columns used for approximation. Here, we develop an efficient SSL algorithm based on multivariate function Taylor Series Expansion (TSE) theory. Our algorithm is based on the Local and Global Consistency SSL algorithm (Zhou, Bousquet et al. 2004). We first use Taylor expansion to the Gaussian kernel function and then we use the Woodbury formula to convert a large matrix inversion problem to a small matrix inversion problem. Experimental comparison of these algorithms will be made in the next section.

3.1 Approximation of Adjacency Matrix Using Multivariate Taylor Expansion

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ be a matrix with each column being a sample $\mathbf{x}_i \in \mathbb{R}^d, i = 1..n$. The edge weight between vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ can be written as

$$w_{ij} = \lambda_i \lambda_j \exp\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\sigma^2}\right) \tag{3}$$

where $\lambda_k = \exp\left(-\frac{\|\mathbf{x}_k\|^2}{2\sigma^2}\right)$ is a scalar. So the adjacency matrix can be rewritten as

$$\mathbf{W} = \mathbf{\Lambda E \Lambda} \tag{4}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\Lambda_{ii} = \lambda_i$ and $\mathbf{E}$ is a matrix with 0s on diagonal and with $e_{ij} = \exp\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\sigma^2}\right)$ on off diagonal. Recall that the Taylor series expansion for a multivariate function $f(\mathbf{x})$ is

$$\begin{aligned} f(\mathbf{x}) = & f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0) + \\ & \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \mathbf{H} (\mathbf{x} - \mathbf{x}_0) + O(\|\mathbf{x} - \mathbf{x}_0\|^2) \end{aligned} \tag{5}$$

where gradient $\nabla f(\mathbf{x}_0) = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0}$ and Hessian $\mathbf{H} = \left.\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T}\right|_{\mathbf{x}=\mathbf{x}_0}$.

So the gradient and Hessian functions of $f(\mathbf{x}) = \exp\left(\frac{\mathbf{x}^T \mathbf{x}}{\sigma^2}\right)$ are

$$\begin{aligned} \nabla f &= \frac{2\mathbf{x}}{\sigma^2} \exp\left(\frac{\mathbf{x}^T \mathbf{x}}{\sigma^2}\right) \\ \mathbf{H} &= \exp\left(\frac{\mathbf{x}^T \mathbf{x}}{\sigma^2}\right)\left(\frac{2}{\sigma^2}\mathbf{I} + \frac{4\mathbf{x}^T \mathbf{x}}{\sigma^4}\right) \end{aligned} \tag{6}$$

$\mathbf{I}$ is the identity matrix. Thus $f(\mathbf{x})$ can be expanded at origin as

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{0}) + \nabla f(\mathbf{0})^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} + O(\|\mathbf{x}\|^2) \\ &= 1 + \frac{\mathbf{x}^T \mathbf{x}}{\sigma^2} + O(\|\mathbf{x}\|^2) \end{aligned} \tag{7}$$

If σ is large enough comparing with $\mathbf{x}^T \mathbf{x}$, then the higher order terms in equation (7) can be ignored and $f(\mathbf{x})$ can be well approximated by the first three terms, i.e.

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{0}) + \nabla f(\mathbf{0})^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} \\ &\approx 1 + \frac{\mathbf{x}^T \mathbf{x}}{\sigma^2} \end{aligned} \tag{8}$$

Plug equation (8) into equation (4), Then approximation of the adjacency matrix is

$$\tilde{\mathbf{W}} = \mathbf{\Lambda}\left(\mathbf{e}\mathbf{e}^T - \mathbf{I} + \frac{1}{\sigma^2}(\mathbf{X}^T \mathbf{X} - \tilde{\mathbf{\Lambda}})\right)\mathbf{\Lambda}^T \tag{9}$$

where matrix $\tilde{\mathbf{\Lambda}}$ is a diagonal matrix with diagonal element $\tilde{\Lambda}_{ii} = \|\mathbf{x}_i\|^2$, and vector $\mathbf{e} = (1,1,...,1)^T$. Furthermore, the volume of vertex *i* is

$$vol(i) = \lambda_i C - \lambda_i^2 + \frac{\lambda_i}{\sigma^2}(\mathbf{x}_i^T \mathbf{y} - \lambda_i \tilde{\Lambda}_{ii}) \tag{10}$$

where $C = \sum_{j=1}^{n} \lambda_j, \mathbf{y} = \sum_{j=1}^{n} \lambda_j \mathbf{x}_j$. Thus the approximation of the normalized adjacency matrix is

$$\begin{aligned} \mathbf{S} &= \mathbf{D}^{-1/2} \tilde{\mathbf{W}} \mathbf{D}^{-1/2} \\ &= \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} - \mathbf{T} \end{aligned} \tag{11}$$

where $\mathbf{T} = \mathbf{D}^{-1/2} \mathbf{\Lambda} \mathbf{\Lambda} \mathbf{D}^{-1/2} + (\mathbf{D}^{-1/2} \mathbf{\Lambda} \tilde{\mathbf{\Lambda}} \mathbf{\Lambda} \mathbf{D}^{-1/2}) / \sigma^2$, $\boldsymbol{\alpha} = \mathbf{D}^{-1/2} \mathbf{\Lambda} \mathbf{e}$ and $\mathbf{H} = \mathbf{X} \mathbf{\Lambda} \mathbf{D}^{-1/2}$. It is worth noting that there are only *n* elements for vector $\boldsymbol{\alpha}$ or diagonal matrix $\mathbf{T}$, and $\mathbf{H}$ is same size as original data matrix. Thus the total space for storing the normalized adjacency matrix reduces tremendously.

3.2 Using Woodbury Formula for Large Matrix Inversion

In equation (11) the spatial complexity is reduced to linear order, but inverting matrix $\mathbf{I} - \gamma \mathbf{S}$ (which is an *n*-by-*n* matrix) is

still time consuming. But if we define a $n\times(d+1)$ matrix $\mathbf{M}=[\boldsymbol{\alpha} \quad \frac{1}{\sigma}\mathbf{H}^T]$, then the normalized adjacency matrix is

$$\tilde{\mathbf{S}}=\mathbf{M}\mathbf{M}^T-\mathbf{T} \tag{12}$$

Using the Woodbury formula (Horn and Johnson 1990) $(\mathbf{AB}+\mathbf{C})^{-1}=\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{A}(\mathbf{I}+\mathbf{B}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{B}\mathbf{C}^{-1}$, then

$$\begin{aligned}(\mathbf{I}-\gamma\tilde{\mathbf{S}})^{-1}&=(\mathbf{I}_n-\gamma\mathbf{M}\mathbf{M}^T+\gamma\mathbf{T})^{-1}\\&=\mathbf{K}^{-1}+\gamma\mathbf{K}^{-1}\mathbf{M}(\mathbf{I}_{d+1}-\gamma\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{K}^{-1}\end{aligned} \tag{13}$$

where $\mathbf{K}=\mathbf{I}+\gamma\mathbf{T}$ is a diagonal matrix, $\mathbf{I}_k$ is the a $k\times k$ identity matrix. Equation (13) only concerns with a diagonal matrix inversion and a $(d+1)\times(d+1)$ matrix inversion, where d is the feature length, usually much smaller than sample number. Thus the final optimal solution in equation (2) is

$$\mathbf{F}^*=\mathbf{K}^{-1}\mathbf{Y}+\mathbf{G}(\gamma\mathbf{I}-\gamma^2\mathbf{M}^T\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{Y}) \tag{14}$$

where $\mathbf{G}=\mathbf{K}^{-1}\mathbf{M}$.

3.3 Complexity analysis

Inverting the diagonal matrix $\mathbf{K}$ is to take reciprocal of its diagonal elements, thus this can be done quickly. The dimension of matrices $\mathbf{M}$ and $\mathbf{G}$ are both $n\times(d+1)$. So $\gamma\mathbf{I}-\gamma^2\mathbf{M}^T\mathbf{G}$ is a $(d+1)\times(d+1)$ matrix, where d is the feature length (e.g. $d=6$ for TM image). Matrix $\mathbf{G}^T\mathbf{Y}$ is a $d\times c$ small matrix, where c is the class number. Therefore, the computation of inverting and multiplication of these matrices is trivial. As it will be seen in the next section, for TM image classification the speed is even faster than SVM, which is generally considered as a fairly efficient supervised classification algorithm. The memory-saving characteristic is also apparent, since the maximum space required for all the computation is as large as two $\mathbf{X}$'s space and several vectors' space.

4. EXPERIMENTAL COMPARISON

We make comparison on computer with 2.13GHZ CPU and Matlab 2011a. For each algorithm, we choose parameters by searching in parameter space and finding optimal one which produces lower error rate. For the AGR, we count running time including the time of finding anchors.

4.1 AVIRIS Image Dataset

The first experimental dataset is a 220-band hyperspectral image: June 12, 1992 AVIRIS image Indian Pine Test Site 3*. The reference ground truth is known. There are 16 land cover types in the ground truth map. We choose 14 of them and discard the other 2 because those 2 types have too few samples for training. We also discard the first 6 bands, which are too noisy and have a very low SNR (Signal Noise Ratio), and use the remaining 214 bands for classification. We generate training samples by randomly selecting k samples as the labelled samples at each time (k=2, 3,…, 50) from each class and mix these 14k labelled samples together to form the training set. Then all the remainder of each class are put together to form the validation set. For each training set, we run the program 5 times and compute the final overall accuracy by averaging. Experimental results are shown in Figure 1 and Figure 2.

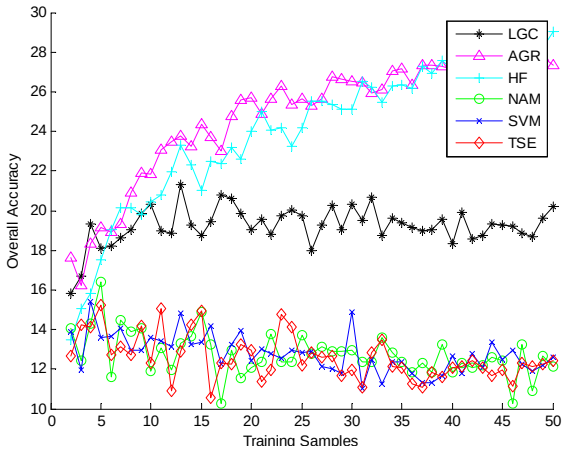

Figure 1. AVIRIS Dataset - Mean Overall Accuracy VS. Training Set Size

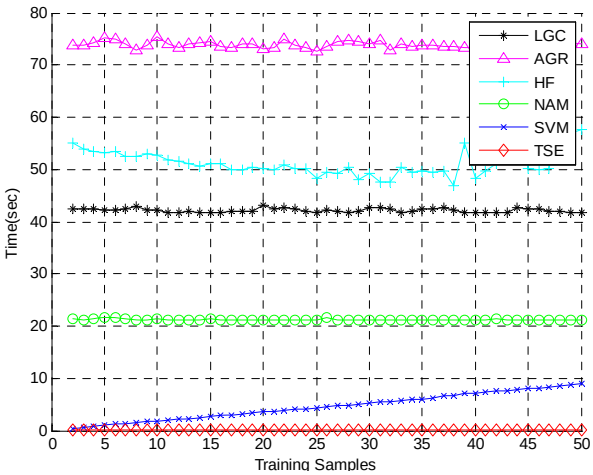

Figure 2. AVIRIS Dataset - Time Cost V.S. Training Set Size

4.2 Landsat Image Dataset

The second experimental dataset is made up of a Landsat-7 Enhanced Thematic Mapper (ETM+) image with path 143 and row 029 acquired in Sep 01, 1999. This region, located in northwest China, is known as the ancient Silk Road, which is the gateway from East Asia to West Asia and Europe. On the north of this region there is the second largest desert in China, Gurbantunggut Desert. This region is on the transition from plain to dessert and contains land cover types varying greatly, and thus it is suitable to test the generalization ability of semi-

*https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html

supervised classifier. A subregion of 660×660 pixels has been considered and its pseudo colour image is shown in figure 3. Four typical land cover classes (i.e. water represented by dark blue pixels, farm represented by red pixels, sand represented by light khaki and residential area by other dark pixels) are defined and their corresponding labelled sample sets are manually determined from ground reference data. The capacities of these labelled sample sets are 1734, 1729, 1087 and 1011, respectively. Bands {1, 2, 3, 4, 5, 7} of the ETM+ imagery are used. In all experiments, the training samples for each land cover type are selected randomly from their corresponding labelled samples sets and the rest of the labelled samples are mixed together as the validation set. In order to compare the effectiveness of classification, we design experiments with training samples varying 1 to 30, i.e. {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30} for each land cover class. So the training sets size are {4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 60, 80, 120} and the validation sets size are {5557, 5553, 5549, 5545, 5541, 5537, 5533, 5529, 5525, 5521, 5501, 5481, 5441}. Classification accuracy is evaluated on these validation sets in the assumption that all the samples' labels are unavailable. Experimental results are shown in figure 4 to figure 6.

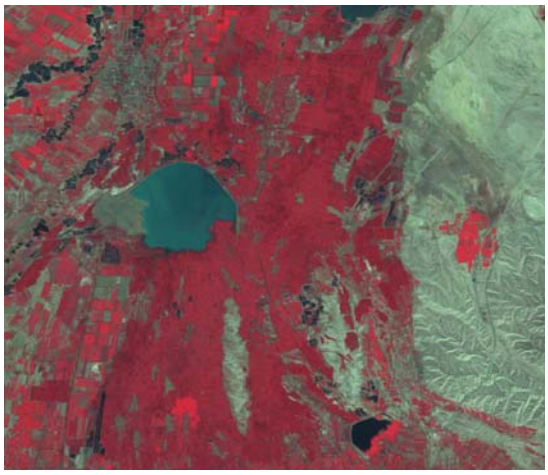

Figure 3. ETM+ Dataset - Pseudo Color Image of The Study Area

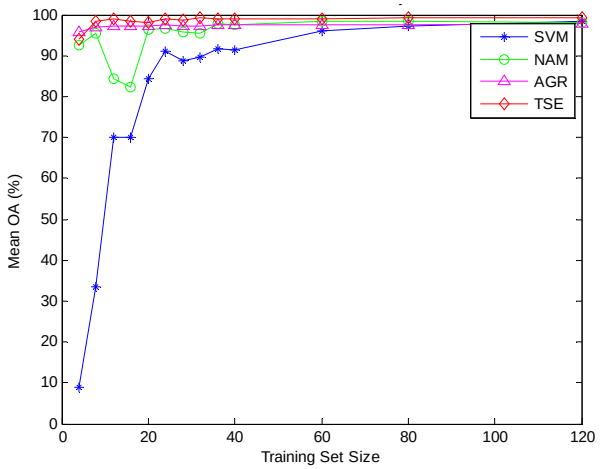

Figure 4. ETM+ Dataset - Mean Overall Accuracy VS. Training Set Size

Figure 4 is the mean overall accuracy vs. training set size and figure 5 is the Standard Variance (StdV) of overall accuracy vs. training set size. From these figures we can infer that SVM have content result (i.e. mean OA ≥ 80% and StdV of OA ≤ 5%) only when the training set size is larger than 20, while the semi-supervised algorithms remain rather stable performance even only when 1 training sample for each class is used. It is worth noting that TSE achieves the best OA and meanwhile keeps a rather lower deviation.

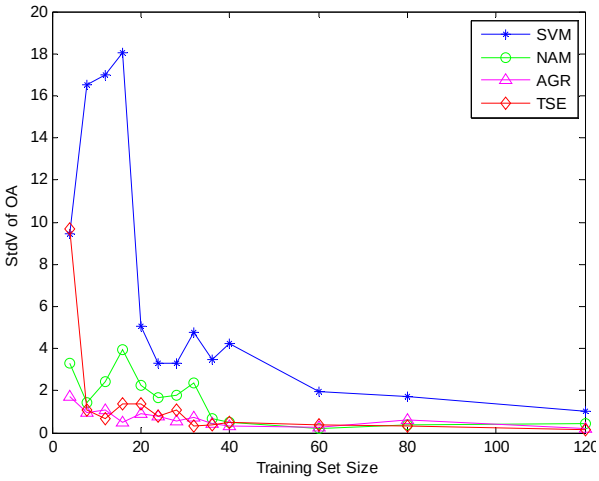

Figure 5. ETM+ Dataset - Standard Variance of Overall Accuracy VS. Training Set Size

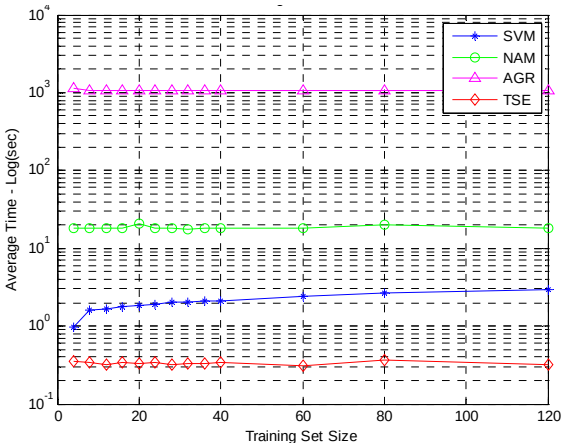

Figure 6. ETM+ Dataset - Average Time Cost V.S. Training Set Size

Figure 6 is the time cost of each classifier vs. training set size, with a logarithmic vertical axis. As expected, the time cost of SVM grows linearly with training set size, and the time cost of the other three semi-supervised learning algorithms remain constant. This can be answered by the fact that as training set grows, the number of support vectors also increases and as a result each test sample needs to compare more times to determine its class label. Thus the accumulated time cost for all the test samples may considerably increase. On the other hand, semi-supervised learning classifiers have time cost only related with the graph on which it learns, so even though the training set size grows, the test samples are unchanged and thus the

graph size remains the same, and therefore the time cost is nearly constant. As is shown in the figure, among the three semi-supervised learning classifiers, TSE has a lower time cost than SVM, which has been considered to be a rather efficient classifier in literature.

5. CONCLUSIONS

We evaluate several standard graph based SSL algorithms in the context of the remote sensing field. To overcome the polynomial complexity of graph based SSL, we also propose an efficient SSL algorithm which has linear complexity in both time and space. To achieve this we first use multivariate Taylor series expansion and then adopt the Woodbury formula for matrix inversion. As it demonstrated, in general, when only a few labelled samples are available, SSL performs no worse (often much better) than traditional supervised methods like SVM, KNN. In experiments we also noted that SSL tends to be not only less sensitive to number of training samples, but also less sensitive to the labelled sample's location on the class distribution. This may be due to the fact that SSL can learn on manifold (Belkin and Niyogi 2004). These properties of SSL have their potential value in the remote sensing field, because in many cases collecting sufficient high quality labelled samples is not a trivial task.

ACKNOWLEDGEMENT

This work is supported by the International Cooperation Project of Ministry of Science and Technology of China with ID: 2009DFA12870 and National Science Foundation No. 61105001.

6. REFERENCES

Belkin, M. and P. Niyogi ,2004. Semi-supervised learning on Riemannian manifolds. Machine Learning, 56(1), pp.209-239.

Belkin, M., P. Niyogi, et al., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research, 7, pp.2399-2434.

Bradley, B. A., 2009. Accuracy assessment of mixed land cover using a GIS-designed sampling scheme. International Journal of Remote Sensing, 30(13), pp.3515-3529.

Bruzzone, L., M. Chi, et al., 2006. A novel transductive SVM for semisupervised classification of remote-sensing images. Geoscience and Remote Sensing, IEEE Transactions on, 44(11), pp.3363-3373.

Camps-Valls, G., T. Bandos Marsheva, et al., 2007. Semi-supervised graph-based hyperspectral image classification. Geoscience and Remote Sensing, IEEE Transactions on, 45(10), pp.3044-3054.

Chapelle, O., B. Schölkopf, et al., 2006. Semi-supervised learning. MIT press Cambridge, MA:.

Chung, F. R. K., 1997. Spectral graph theory, American Mathematical Society.

Fergus, R., Y. Weiss, et al., 2009. Semi-supervised learning in gigantic image collections. Advances in Neural Information Processing Systems, 1.

Friedl, M., D. McIver, et al., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote Sensing of Environment, 83(1-2), pp. 287-302.

Horn, R. A. and C. R. Johnson, 1990. Matrix analysis, Cambridge Univ Pr.

Liu, W., J. He, et al., 2010. Large graph construction for scalable semi-supervised learning, ICML.

Lu, D., P. Mausel, et al., 2004. Change detection techniques. International Journal of Remote Sensing, 25(12), pp.2365-2401.

Marconcini, M., G. Camps-Valls, et al., 2009. A composite semisupervised SVM for classification of hyperspectral images. Geoscience and Remote Sensing Letters, IEEE, 6(2), pp. 234-238.

Pauleit, S. and F. Duhme, 2000. Assessing the environmental performance of land cover types for urban planning. Landscape and Urban Planning, 52(1), pp.1-20.

Powell, R., N. Matzke, et al., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sensing of Environment, 90(2), pp.221-234.

Radke, R. J., S. Andra, et al., 2005. Image change detection algorithms: a systematic survey. Image Processing, IEEE Transactions on,14(3), pp.294-307.

Shi, J. and J. Malik , 2000. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(8), pp.888-905.

Sobrino, J. and N. Raissouni, 2000. Toward remote sensing methods for land cover dynamic monitoring: application to Morocco. International Journal of Remote Sensing, 21(2), pp.353-366.

Zhou, D., O. Bousquet, et al., 2004. Learning with local and global consistency. NIPS.

Zhu, X., 2006. Semi-supervised learning literature survey. Computer Science, University of Wisconsin-Madison.

Zhu, X., Z. Ghahramani, et al., 2003. Semi-supervised learning using gaussian fields and harmonic functions. ICML.

Zhu, X. and A. B. Goldberg, 2009. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3(1), pp.1-130.