

Semi-Supervised Graph Embedding Scheme with Active Learning (SSGEAL): Classifying High Dimensional Biomedical Data

George Lee and Anant Madabhushi

Rutgers, The State University of New Jersey,
Department of Biomedical Engineering
Piscataway, NJ 08854 USA

geolee@eden.rutgers.edu, anantm@rci.rutgers.edu

Abstract. In this paper, we present a new dimensionality reduction (DR) method (SSGEAL) which integrates Graph Embedding (GE) with semi-supervised and active learning to provide a low dimensional data representation that allows for better class separation. Unsupervised DR methods such as Principal Component Analysis and GE have previously been applied to the classification of high dimensional biomedical datasets (e.g. DNA microarrays and digitized histopathology) in the reduced dimensional space. However, these methods do not incorporate class label information, often leading to embeddings with significant overlap between the data classes. Semi-supervised dimensionality reduction (SSDR) methods have recently been proposed which utilize both labeled and unlabeled instances for learning the optimal low dimensional embedding. However, in several problems involving biomedical data, obtaining class labels may be difficult and/or expensive. SSGEAL utilizes labels from instances, identified as “hard to classify” by a support vector machine based active learning algorithm, to drive an updated SSDR scheme while reducing labeling cost. Real world biomedical data from 7 gene expression studies and 3900 digitized images of prostate cancer needle biopsies were used to show the superior performance of SSGEAL compared to both GE and SSAGE (a recently popular SSDR method) in terms of both the Silhouette Index (SI) (SI = 0.35 for GE, SI = 0.31 for SSAGE, and SI = 0.50 for SSGEAL) and the Area Under the Receiver Operating Characteristic Curve (AUC) for a Random Forest classifier (AUC = 0.85 for GE, AUC = 0.93 for SSAGE, AUC = 0.94 for SSGEAL).

1 Introduction

Dimensionality reduction (DR) is useful for extracting a few relatively simple patterns from more complex data. For very high dimensional data, such as gene expression, the original feature space could potentially span up to tens of thousands of features. This makes it difficult to build generalizable predictors on account of the curse of dimensionality problem [1], where the feature space is much larger than the number of samples available for classifier training. Therefore, DR methods are often utilized as a precursor to classification. Predictors

can then be trained on low dimensional embedded features, resulting in improved classification accuracy while also allowing researchers to visualize and interpret relationships between data points [1].

Most commonly used DR methods, such as Principal Component Analysis (PCA) [2], Graph Embedding [3], or Manifold Learning [2] schemes are unsupervised, meaning they do not take into account class label information. These methods essentially use cost functions assuming that the best features lie in a subspace of the original high dimensional space where most of the variance in the data is centered. Supervised DR methods such as linear discriminant analysis (LDA) [1] employ cost functions where class labels are incorporated to help separate known classes in a low dimensional embedding.

LDA is one of the most popular supervised DR methods; however it does not consider unlabeled instances [1, 4]. Blum et al. [5] suggested that incorporating unlabeled samples in addition to labeled samples can significantly improve classification results. Subsequently, many new DR methods employ semi-supervised (SS) or weakly labeled learning techniques which incorporate the use of both labeled and unlabeled data [4, 6–9]. These SSSDR schemes use labeled information in the construction of a pairwise similarity matrix, where the individual cells are assigned weights based on class and feature-based similarity between sample pairs. These weights can then be used to create a low dimensional mapping by solving a simple eigen-problem, the hypothesis being that embeddings explicitly employing label information result in greater class separation in the reduced dimensional space.

Active Learning (AL) algorithms have been utilized to intelligently identify hard to classify instances. By querying labels for only hard to classify instances, and using them to train a classifier, the resulting classifier has higher classification accuracy compared to random learning, assuming the same number of queries are used for classifier training [10, 11]. In practice, obtaining labels for biomedical data is often expensive. For example, in the case of digital pathology applications, disease extent can only be reliably annotated by an expert pathologist. By employing AL, the predictive model is (a) cheaper to train and (b) yields a superior decision boundary for improved discrimination between object classes with fewer labeled instances.

In this paper we present Semi-Supervised Graph Embedding with Active Learning (SSGEAL), a new DR scheme for analysis and classification of high dimensional, weakly labeled data. SSGEAL identifies the most difficult to classify samples via a support vector machine based active learning scheme, which is then used to drive a semi-supervised graph embedding algorithm. Predictors can then be trained for object classification in the SSGEAL reduced embedding space.

2 Previous Work and Novel Contributions

2.1 Unsupervised Dimensionality Reduction

PCA is the most commonly used unsupervised DR method. However it is essentially a linear DR scheme [2]. Nonlinear dimensionality reduction (NLDR)

methods such as Isomap [2] and Locally Linear Embedding [2], are powerful due to their ability to discover nonlinear relationships between samples. In [1], we found that nonlinear DR schemes outperformed PCA for the problem of classifying high dimensional gene- and protein-expression datasets. However, NLDR schemes are notoriously unstable [1, 2], requiring careful tuning of a neighborhood parameter to generate useful embeddings.

Graph Embedding [3], or Spectral Embedding is an alternative unsupervised NLDR method which does not require adjusting a neighborhood parameter, and has been found to be useful in applications involving classification of DNA microarrays, proteomic spectra, and biomedical imaging [1, 12]. Normalized cuts [3] is one implementation of Graph Embedding, which is widely used in the area of image segmentation. Other versions of graph embedding include Min Cut [5], Average Cut [3], Associative Cut [3], and Constrained Graph Embedding [13].

2.2 Semi-Supervised Dimensionality Reduction

Sugiyama et al. [4] applied SS-learning to Fisher's discriminant analysis in order to find projections that maximize class separation. Yang et al. [8] similarly applied SS-learning toward manifold learning methods. Sun et al. [9] implemented a SS version of PCA by exploiting between-class and within-class scatter matrices. SSAGE [6] is a SS method for spectral clustering which utilizes weights to simultaneously attract within-class samples and repel between-class samples given a neighborhood constraint. However, these embeddings often contain unnatural, contrived clusters on account of labeled samples. Zhang [7] uses a similar approach to SSDLR, but without utilizing neighborhood constraints.

2.3 Active Learning

Previous AL methods have looked at the variance of sample classes to identify difficult to classify instances [14]. The Query by Committee approach [10] uses disagreement across several weak classifiers to identify hard to classify samples. In [15], a geometrically based AL approach utilized support vector machines (SVMs) to identify confounding samples as those that lay closest to the decision hyperplane. SVM-based AL has previously been applied successfully to the problem of classifying gene expression data [11]. Additionally, a clear and easily interpretable rationale for choice of sample selection exists. All these methods however have typically been applied to improving classification and not embedding quality per se [10, 14].

2.4 Novel Contributions and Significance of SSGEAL

The primary contribution of this paper is that it merges two powerful schemes - SSDLR with Active Learning - for generating improved low dimensional embedding representations, which allows for greater class separation.

Figure 1 illustrates how Graph Embedding (GE) can be improved with SS-learning (SSAGE), and even further using AL (SSGEAL). In Figure 1(a), a simple RGB image consisting of ball and background pixels is shown. Following the addition of Gaussian noise, each pixel in Figure 1a is plotted in a 3D RGB space (Figure 1(e)). Subsequently, we reduce the 3D RGB space into a 2D embedding via GE (Figure 1(f)), SSAGE (Figure 1(g)), and SSGEAL (Figure 1(h)). Figures 1(b), 1(c), and 1(d) represent a pixel-wise binary classification into foreground (ball) and background classes via GE, SSAGE, and SSGEAL, respectively. These were obtained via replicated k-means clustering on the corresponding DR embeddings, as shown in Figures 1(f), 1(g), and 1(h).

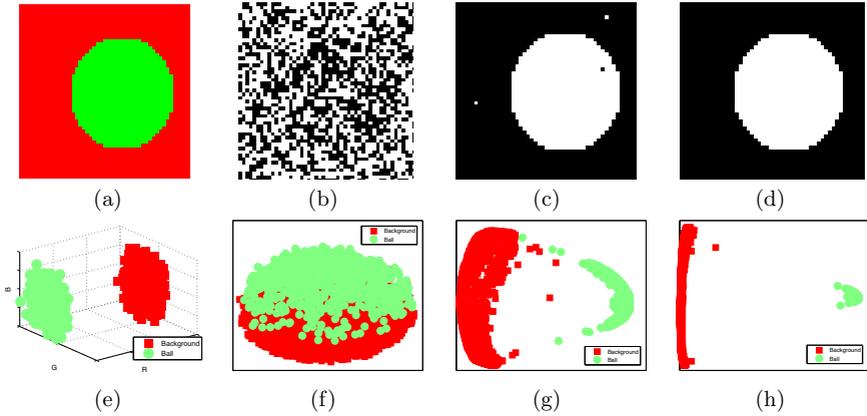


Fig. 1. (a) RGB image containing ball against colored background pixels. (e) image pixels plotted in 3D RGB space. The binary classifications (b-d) reflect the corresponding quality of embeddings obtained via DR methods (b) GE, (c) SSAGE, and (d) SSGEAL. These were obtained via replicated k-means clustering on the reduced embeddings by (f) GE, (g) SSAGE, and (h) SSGEAL, respectively.

Table 1. Commonly used notation in this paper

<i>Symbol</i>	<i>Description</i>
X	Set containing N samples
$\mathbf{x}_i, \mathbf{x}_j$	Sample vector $\mathbf{x}_i, \mathbf{x}_j \in X, i, j \in \{1, 2, \dots, N\}, \mathbf{x} \in \mathbb{R}^n$
n	Number of features used to describe \mathbf{x}_i
W	Dissimilarity matrix
$Y(\mathbf{x}_i)$	Labels for samples $\mathbf{x}_i, Y(\mathbf{x}_i) \in \{+1, -1\}$
$Z(X, Y(X_{Tr}))$	Embedding Z constructed using data X and label set $Y(X_{Tr})$.
X_{Tr}	Set of labeled training samples $\mathbf{x}_i \in X_{Tr}$
X_{Ts}	Set of unlabeled testing samples $X_{Ts} \subset X$
X_a	Set of ambiguous samples $X_a \subset X_{Ts}$
δ	Distance to decision hyperplane F in SVM-based AL

3 Review of SSSR and Active Learning Methods

3.1 Graph Embedding (GE)

To obtain low dimensional embedding Z , Graph Embedding [3] utilizes pairwise similarities between objects \mathbf{x}_i and $\mathbf{x}_j \in X$ to construct $N \times N$ weighted graph

$$W(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}, \quad (1)$$

where $\sigma = 1$. $\tilde{W}(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{ii} W(\mathbf{x}_{ii}, \mathbf{x}_j) \times \sum_{jj} W(\mathbf{x}_i, \mathbf{x}_{jj}))^{-1} W(\mathbf{x}_i, \mathbf{x}_j)$ is then used to solve the eigenvalue problem $(D - \tilde{W})\mathbf{z} = \lambda D\mathbf{z}$, where D is a diagonal matrix containing the trace of \tilde{W} , and \mathbf{z}_k are the eigenvectors. Embedding Z is formed by taking the most dominant eigenvectors \mathbf{z}_k corresponding to the k smallest eigenvalues λ_k , where k is the dimensionality of Z . In this implementation, Graph Embedding does not consider labeled information.

3.2 Semi-Supervised Agglomerative Graph Embedding (SSAGE)

By using known label information, Zhao [6], describes a method for SSSR where the similarity weights for GE are adjusted such that Equation 1 is replaced by

$$W(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} (e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}})(1 + e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}), & \text{if } Y(\mathbf{x}_i) = Y(\mathbf{x}_j) \\ (e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}})(1 - e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}), & \text{if } Y(\mathbf{x}_i) \neq Y(\mathbf{x}_j) \\ e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}, & \text{otherwise} \end{cases} \quad (2)$$

In contrast to simple GE, in SSAGE, known labeled samples are mapped to be closer in the embedding space Z if both samples \mathbf{x}_i and \mathbf{x}_j are of the same class $Y(\mathbf{x}_i) = Y(\mathbf{x}_j)$, and further apart if both samples are of different classes.

3.3 SVM-Based Active Learning to Identifying Ambiguous Samples

A labeled set X_{Tr} is first used to train the SVM. SVMs [16] project the input training data onto a high-dimensional space using the kernel $\Pi(\mathbf{x}_i, \mathbf{x}_j)$. A linear

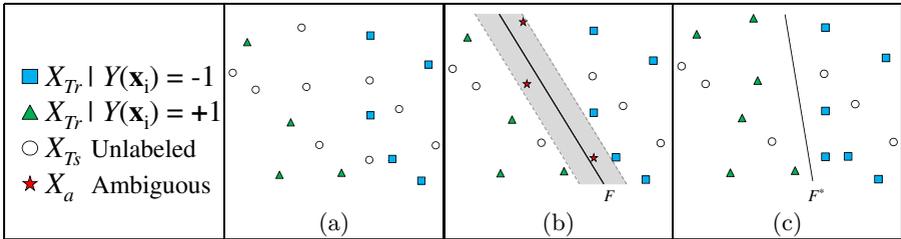


Fig. 2. (a) Labeled samples $\mathbf{x}_i \in X_{Tr}$ are used to train an SVM model F . (b) Unlabeled samples $\mathbf{x}_i \in X_{Ts}$ found to be mapped closest to the model hyperplane F are included into set X_a . (c) Labels $Y(\mathbf{x}_i \in X_a)$ are queried and used to improve the new SVM model F^* , yielding a better predictor compared to F .

kernel, defined as $\Pi(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$, can then be used to maximize the margins so as to decrease prediction risk. A decision boundary F is created in the trained feature space by maximizing the margin between classes. Unlabeled instances $\mathbf{x}_i \in X_{T_s}$ are mapped into the same feature space (Figure 2(a)).

However, instead of classifying X_{T_s} , we use boundary F to find ambiguous samples $\mathbf{x}_i \in X_a$ via measure δ , defined as the relative distance to hyperplane F . Samples $\mathbf{x}_i \in X_{T_s}$ of shortest δ represent the most ambiguous samples and are assigned to set X_a (Figure 2(b)). Labels for X_a are queried and these ambiguous samples are added to the subsequent training set $X_{T_r} = [X_{T_r}, X_a]$. Learning via the updated labels $Y(X_{T_r})$ results in improved class separation (Figure 2(c)).

4 Semi-Supervised Graph Embedding with Active Learning (SSGEAL)

4.1 Initialization with Initial Embedding Z_0

The schema for SSGEAL is illustrated via the flowchart in Figure 3. Our initialization comprises of creating an initial embedding Z_0 and defining the initial training X_{T_r} for our active learning scheme within Z_0 . Given data set X , we use Graph Embedding as illustrated in Section 3.1 to obtain our initial embedding $Z_0(X) = [\mathbf{z}_1, \dots, \mathbf{z}_k]$, or simply Z_0 .

4.2 Active Learning to Identify Ambiguous Samples X_a

SVM-based active learning (see Section 3.3) is used to identify ambiguous samples $\mathbf{x}_i \in X_a$ in embedding Z_q , where q represents the specific iteration of an embedding Z . Initial labeled training samples X_{T_r} for AL are selected randomly from X . We begin by training an SVM using $Z_0(X_{T_r})$ and $Y(X_{T_r})$ to create model F . $\delta(X_{T_s})$ can be found using F , where the smallest $\delta(X_{T_s})$ are selected and assigned to set X_a . $Y(X_a)$ is revealed and X_a is added to the training set X_{T_r} , such that $X_{T_r} = [X_{T_r}, X_a]$.

4.3 Semi-Supervised Graph Embedding Z_q Using Updated Labels

We utilize an updated version of Zhao’s SSAGE method [6] to map a modified similarity matrix W into Z using the GE framework discussed in Section 3.1. This weighting only takes into account samples which are of the same class, using a gravitation constant $G > 1$ to attract same-class samples closer. Weights are adjusted such that Equation 1 is replaced by

$$W(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} G \times e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}, & \text{if } Y(\mathbf{x}_i) = Y(\mathbf{x}_j) \\ e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}, & \text{otherwise} \end{cases} \quad (3)$$

Unlike the Zhao [6] and Zhang [7] implementations, instances from different classes are not explicitly weighted to force them farther apart in SSGEAL. The

rationale for this is that for biomedical data, certain instances within one class may share several traits with another class. For instance, premalignant lesions while technically benign, share several hallmarks of malignant tumors. Artificially forcing instances from different classes farther apart could result in a pre-malignant lesion being mapped far apart from the cancer class, rather than in an intermediate class between benign and malignant.

Labels $Y(X_{Tr})$ from the updated training set and current embedding Z_q are used to create embedding Z_{q+1} . The new embedding $Z_{q+1}(Z_q, Y(X_{Tr}))$, or simply Z_{q+1} , is constructed using the current embedding Z_q and the exposed label set $Y(X_{Tr})$. The process of obtaining new labels from AL and creating semi-supervised embeddings continues until the stopping criterion is met.

4.4 Stopping Criterion Using Silhouette Index

The stopping criterion is set using the Silhouette Index (ϕ^{SI}) [17] of the revealed labels. ϕ^{SI} is a cluster validity measure which captures the intra-cluster compactness $A_i = \sum_{j, Y(\mathbf{x}_j)=Y(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{x}_j\|_2$, which represents the average distance of a point \mathbf{x}_i from other points X_j of the same class, while also taking into account inter-cluster separation $B_i = \sum_{j, Y(\mathbf{x}_j) \neq Y(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{x}_j\|_2$, the minimum of the average distances of a point \mathbf{x}_i from other instances in different classes. Thus, the formulation for Silhouette Index is shown as

$$\phi^{SI} = \sum_i^N \frac{B_i - A_i}{\max[A_i, B_i]}. \quad (4)$$

ϕ^{SI} ranges from -1 to 1, where -1 is the worst, and 1 is the best possible clustering. When the change in ϕ^{SI} falls below threshold θ , such that $|\phi_{q+1}^{SI} - \phi_q^{SI}| < \theta$, the algorithm stops. The algorithm for SSGEAL is presented below.

Algorithm *SSGEAL*

Input: $X, Y(X_{Tr})$ $\theta, \delta, q = 0$

Output: Z_f

begin

0. Build initial embedding $Z_0(X)$
1. **while** $|\phi_{q+1}^{SI} - \phi_q^{SI}| < \theta$
2. Train SVM model F using $X_{Tr}, Y(X_{Tr})$
3. Identify $\mathbf{x}_i \in X_a$ using measure δ
4. Update $X_{Tr} = [X_{Tr}, X_a]$
5. Update embedding $Z_{q+1}(Z_q, Y(X_{Tr}))$ via Equation 3
6. Compute ϕ_q^{SI} using Equation 4.
7. $q = q + 1$
8. **endwhile**
9. **return** Z_f

end

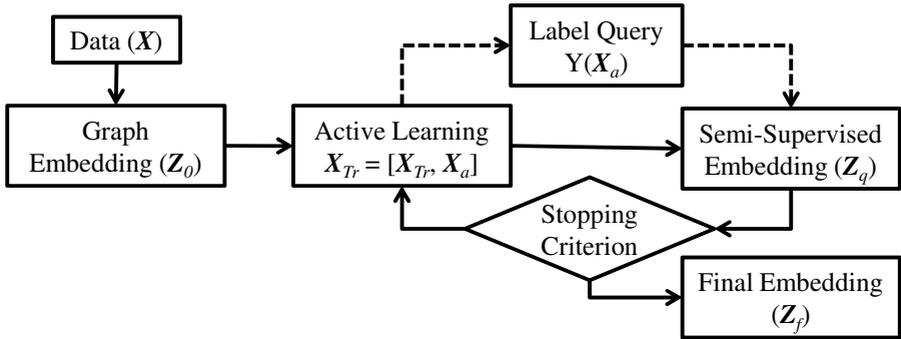


Fig. 3. Flowchart of SSGEAL

5 Experimental Results and Discussion

5.1 Experiments and Evaluation

Datasets. Table 2 provides an overview of the 7 publically available gene expression and digitized prostate biopsy images used to test SSGEAL.¹ For the gene expression datasets, no preprocessing or normalization of any kind was performed prior to DR. For the Prostate Histopathology dataset, a set of 14 pixel-wise features were extracted, including first-order statistical, second-order co-occurrence, and steerable Gabor wavelet features [10, 18] from the images, digitized at 40x magnification. The images are then broken into 30 x 30 pixel regions, each quantified by averaging the feature values in the region. We randomly selected 3900 non-overlapping patches from within the cancer and non-cancer regions (manually annotated by an expert pathologist) for purposes of evaluation.

Table 2. Datasets used in our experiments

	Datasets	Description
Gene Expression	Prostate Cancer	25 Tumor, 9 Normal, 12600 genes
	Colon Cancer	22 Tumor, 40 Normal, 2000 genes
	Lung Cancer	15 MPM, 134 ADCA, 12533 genes
	ALL / AML	20 ALL, 14 AML, 7129 genes
	DLBCL Tumor	58 Tumor, 19 Normal, 6817 genes
	Lung Cancer(Mich)	86 Tumor, 10 Normal, 7129 genes
	Breast Cancer	10 Tumor, 20 Normal, 54675 genes
Imaging	Prostate Histopathology	1950 cancer regions, 1950 benign regions, 14 image textural descriptors

¹ Gene expression datasets were obtained from the Biomedical Kent-Ridge Repositories at <http://sdmc.lit.org.sg/GEDatasets/Datasets> and <http://sdmc.i2r.a-star.edu.sg/rp>

Table 3. $\mu(\phi^{SI})$, $\mu(\phi^{AUC})$, $\sigma(\phi^{SI})$, and $\sigma(\phi^{AUC})$ across 10 runs using different X_{Tr} for GE, SSAGE and SSGEAL. The high mean performance and low standard deviation of these statistics over 10 runs of SSGEAL on 8 datasets demonstrates the robustness of the algorithm regardless of initial training set X_{Tr} . Best values are shown in **bold**. For a majority of the cases, SSGEAL is shown to perform the best.

	Datasets	Silhouette Index			Random Forest AUC		
		GE	SSAGE	SSGEAL	GE	SSAGE	SSGEAL
Gene Expression	Prostate Cancer	0.54	0.29±0.10	0.66±0.01	1.00	0.98±0.04	1.00±0.00
	Colon Cancer	0.02	0.16±0.01	0.43±0.04	0.73	0.92±0.03	0.95±0.05
	Lung Cancer	0.64	0.49±0.06	0.65±0.20	0.49	0.95±0.10	0.96±0.09
	ALL / AML	0.42	0.24±0.04	0.47±0.05	0.95	0.96±0.03	0.97±0.04
	DLBCL Tumor	0.20	0.32±0.10	0.62±0.03	0.75	0.89±0.04	0.95±0.04
	Lung Cancer(Mich)	0.68	0.45±0.02	0.83±0.02	1.00	0.95±0.13	0.99±0.03
	Breast Cancer	0.20	0.19±0.09	0.45±0.08	0.78	0.90±0.05	0.96±0.05
Imaging	Prostate Histopathology	0.35	0.36±0.00	0.35±0.00	0.85	0.93±0.00	0.93±0.00

Experiments. Two DR techniques were employed to compete against our algorithm (SSGEAL): one which does not incorporate labels (GE) and one which utilizes labels (SSAGE). We generated embeddings Z using DR methods GE, SSAGE, and SSGEAL to show that (a) embeddings generated using SSGEAL outperform those generated via GE and SSAGE, (b) steady improvement in both classification accuracy and Silhouette index can be observed via active learning with SSGEAL, and (c) SSGEAL is robust to initial training.

Evaluation Measures. Embeddings were evaluated both qualitatively and quantitatively using ϕ^{SI} (Equation 4) and Area Under the Receiver Operating Characteristic (ROC) Curve for Random Forest Classification ϕ^{AUC} . For ϕ^{SI} , all labels were used. For ϕ^{AUC} , a randomly selected training pool \mathcal{P} consisting of two-thirds of the instances in X was used, with the remaining samples reserved for testing. 50 decision trees were trained using a 50 random subsets each consisting of 2/3 of \mathcal{P} . Predictions on the testing samples were subsequently bagged and used to calculate the ROC curve for assessing classifier performance.

Parameter Settings. For our experiments, 2D embeddings $Z = [\mathbf{z}_1, \mathbf{z}_2]$ are generated for each DR method. In all cases, no neighborhood information was used. For both SSAGE and SSGEAL, we ultimately expose 40% of the labels. For SSGEAL, the gravitation constant G was set to 1.3 and our initial training set X_{Tr} was set at 15% of $Y(X)$, revealing 5% of the labels $Y(X_{T_s})$ at each iteration q until 40% of the labels were revealed.

5.2 Comparing SSGEAL with GE and SSAGE via ϕ^{SI} and ϕ^{AUC}

Table 3 lists the mean and variance of ϕ^{AUC} and ϕ^{SI} values for SSGEAL, GE, and SSAGE, over 8 dataset. The same number of labeled samples (40%) were used for SSAGE and SSGEAL for each data set. To obtain an accurate representation of algorithm performance, we randomly selected 10 training sets X_{Tr} for

10 runs of SSAGE and SSGEAL for the purpose of testing the robustness of the algorithms to initial labeling. Note that GE is an unsupervised method and does not utilize label information, hence there is no standard deviation across multiple runs of GE. 2D embeddings were generated for each set X_{T_r} and evaluated via ϕ^{AUC} and ϕ^{SI} .

For a majority of the datasets, SSGEAL outperforms both GE and SSAGE in terms of ϕ^{SI} ($\mu(\phi^{SI})$ of 0.35 for GE, 0.31 for SSAGE, and 0.50 for SSGEAL) and ϕ^{AUC} ($\mu(\phi^{AUC})$ of 0.85 for GE, 0.93 for SSAGE, and 0.94 for SSGEAL). Furthermore, low standard deviation ($\sigma(\phi^{AUC}), \sigma(\phi^{SI})$) over the 10 runs suggest robustness of SSGEAL to initial X_{T_r} .

Figure 4 shows qualitative illustrations of 2D embeddings for GE and SS-GEAL over different iterations for 3 selected datasets. We can observe greater class separation and cluster tightness with increasing iterations for SSGEAL.

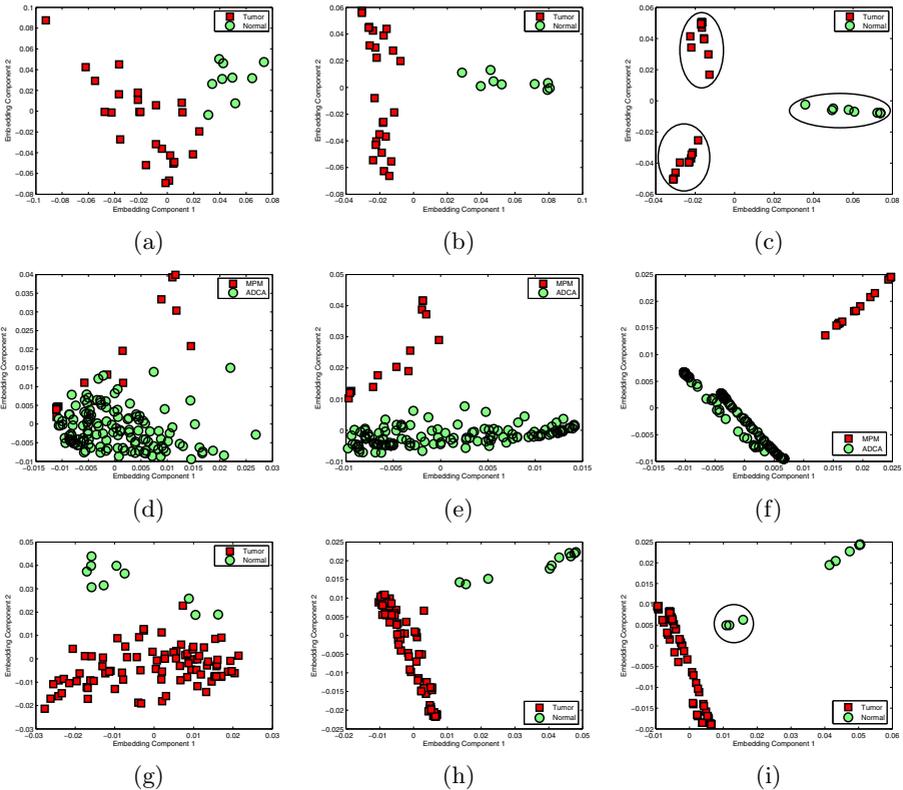


Fig. 4. Scatter plots of the 2 most dominant embedding eigenvectors $\mathbf{z}_1(\mathbf{x}_i)$, $\mathbf{z}_2(\mathbf{x}_i)$ for $\mathbf{x}_i \in X$ are shown for different iterations of SSGEAL (a) Z_0 , (b) Z_2 , and (c) Z_f (the final stable embedding), for the Prostate Cancer dataset. Similarly, the embedding plots are shown for the Lung Cancer dataset for (d) Z_0 , (e) Z_2 , (f) Z_f . Lastly, (g) Z_0 , (h) Z_2 , (i) Z_f are shown for the Lung Cancer(Mich) dataset. Note the manually placed ellipses in (c) and (i) highlight what appear to be novel subclasses.

Figures 4(a), 4(d), and 4(g) show embedding plots of GE (Z_0). An intermediate step of SSGEAL (Z_q) is shown in Figures 4(b), 4(e), and 4(h) and SSGEAL embeddings (Z_f) can be seen in Figures 4(c), 4(f), and 4(i).

6 Concluding Remarks

Semi-Supervised Graph Embedding with Active Learning (SSGEAL) represents the first attempt at incorporating an active learning algorithm into a semi-supervised dimensionality reduction (SSDR) framework. The inclusion of active learning is especially important for problems in biomedical data where class labels are often difficult or expensive to come by. Using 8 real-world gene expression and digital pathology image datasets, we have shown that SSGEAL results in low dimensional embeddings which yield tighter, more separated class clusters and result in greater class discriminability compared to GE and SSAGE, as evaluated via the Silhouette Index and AUC measures. Furthermore, SSGEAL was found to be robust with respect to the choice of initial labeled samples used for initializing the active learning process. SSGEAL does however appear to be sensitive to the value assigned to the gravitation constant G . This parameter may be used to refine the initial graph embedding (Figure 5(a)). For the histology dataset, setting $G = 1.5$ resulted in $\phi^{SI} = 0.39$ and $\phi^{AUC} = 0.94$ for SSGEAL, compared to $\phi^{SI} = 0.36$ and $\phi^{AUC} = 0.93$ for SSAGE (Figure 5). In future work we intend to extensively and quantitatively evaluate the sensitivity of our scheme to neighborhood, gravitation, and stopping parameters.

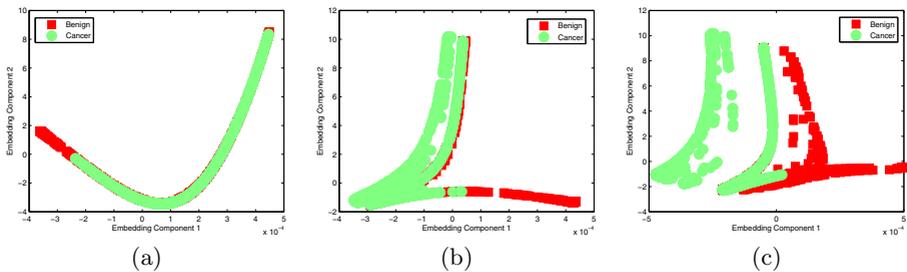


Fig. 5. Scatter plots of the 2 most dominant embedding eigenvectors are shown for the Prostate Histopathology dataset. (b) and (c) show SSGEAL embeddings with gravitation constants $G = 1.3$ and 1.5 respectively, suggesting the utility of G for improving embeddings with large degrees of overlap between the object classes. For comparison, the embedding graph for GE is also shown for this dataset (Figure 5(a)).

Acknowledgments. Funding for this work made possible by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, R21CA127186-01, R03CA128081-01), the Cancer Institute of New Jersey, and Bioimagine Inc.

References

1. Lee, G., Rodriguez, C., Madabhushi, A.: Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE Trans. on Comp. Biol. and Bioinf.* 5(3), 368–384 (2008)
2. van der Maaten, L.J.P., Postma, E.O., van den Herik, H.J.: Dimensionality reduction: A comparative review. Tilburg University Technical Report, TiCC- TR2009–005 (2009)
3. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence.* 22(8), 888–905 (2000)
4. Sugiyama, M., Idé, T., Nakajima, S., Sese, J.: Semi-supervised local fisher discriminant analysis for dimensionality reduction. *Advances in Knowledge Discovery and Data Mining*, 333–344 (2008)
5. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph min-cuts. In: *International Conference on Machine Learning*, pp. 19–26 (2001)
6. Zhao, H.: Combining labeled and unlabeled data with graph embedding. *Neurocomputing* 69(16–18), 2385–2389 (2006)
7. Zhang, D., et al.: Semi-supervised dimensionality reduction. In: *SIAM International Conference on Data Mining* (2007)
8. Yang, X., Fu, H., Zha, H., Barlow, J.: Semi-supervised nonlinear dimensionality reduction. In: *International Conference on Machine Learning*, pp. 1065–1072 (2006)
9. Sun, D., Zhang, D.: A new discriminant principal component analysis method with partial supervision. *Neural Processing Letters* 30, 103–112 (2009)
10. Doyle, S., et al.: A class balanced active learning scheme that accounts for minority class problems: Applications to histopathology. In: *MICCAI* (2009)
11. Liu, Y.: Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inf. Comput. Sci.* 44(6), 1936–1941 (2004)
12. Higgs, B.W., et al.: Spectral embedding finds meaningful (relevant) structure in image and microarray data. *BMC Bioinformatics* 7(74) (2006)
13. He, X., Ji, M., Bao, H.: Graph embedding with constraints. In: *International Joint Conference on Artificial Intelligence*, pp. 1065–1070 (2009)
14. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. *Journal of Artif. Intell. Res.* 4, 129–145 (1996)
15. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 999–1006 (2000)
16. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learning* 20 (1995)
17. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20(1), 53–65 (1987)
18. Doyle, S., Tomaszewski, J., Feldman, M., Madabhushi, A.: Hierarchical boosted bayesian ensemble for prostate cancer detection from digitized histopathology. *IEEE Transactions on Biomedical Engineering* (2010)