An Empirical Comparison of Dimensionality Reduction Methods for Classifying Gene and Protein Expression Datasets

George Lee[1], Carlos Rodriguez[2], and Anant Madabhushi[1]

[1] Rutgers, The State University of New Jersey,
Department of Biomedical Engineering,
Piscataway, NJ 08854 USA
anantm@rci.rutgers.edu
[2] University of Puerto Rico
Mayagez, PR 00681-9000

Abstract. The recent explosion in availability of gene and protein expression data for cancer detection has necessitated the development of sophisticated machine learning tools for high dimensional data analysis. Previous attempts at gene expression analysis have typically used a linear dimensionality reduction method such as Principal Components Analysis (PCA). Linear dimensionality reduction methods do not however account for the inherent nonlinearity within the data. The motivation behind this work is to demonstrate that nonlinear dimensionality reduction methods are more adept at capturing the nonlinearity within the data compared to linear methods, and hence would result in better classification and potentially aid in the visualization and identification of new data classes. Consequently, in this paper, we empirically compare the performance of 3 commonly used linear versus 3 nonlinear dimensionality reduction techniques from the perspective of (a) distinguishing objects belonging to cancer and non-cancer classes and (b) new class discovery in high dimensional gene and protein expression studies for different types of cancer. Quantitative evaluation using a support vector machine and a decision tree classifier revealed statistically significant improvement in classification accuracy by using nonlinear dimensionality reduction methods compared to linear methods.

Keywords: dimensionality reduction, bioinformatics, gene expression, proteomics, classification, prostate cancer, lung cancer, ovarian cancer, principal component analysis, linear discriminant analysis, multidimensional scaling, graph embedding, Isomap, locally linear embedding.

1 Introduction

The information found in gene and protein expression studies provides a means for identifying patients with cancer and hence these studies have emerged as promising techniques for cancer detection [1,2]. A typical gene expression dataset, however, contains information from thousands of genes (features), which are likely

I. Măndoiu and A. Zelikovsky (Eds.): ISBRA 2007, LNBI 4463, pp. 170–181, 2007.

to be significantly greater than the number of patients from whom the data was collected. The relatively small number of patient samples compared to the very large size of the feature space results in the so-called 'curse of dimensionality' problem from a data analysis perspective [3]. Many of the genes within the expression studies may be non-informative or redundant and hence may not contribute very much from a classification perspective [4]. Two common approaches to making the data amenable to classification are (i) feature selection and (ii) dimensionality reduction (DR).

Feature selection refers to the elimination of genes determined as either being highly correlated with other genes or non-informative with respect to distinguishing the data classes [4]. It serves as a direct method for reducing inherent data dimensionality prior to classification by acquiring an optimal subset of genes to maximally separate the data classes. However, since a typical gene microarray records thousands of gene expressions, each associated with a particular gene, the cost of finding an optimal subset from several million possible combinations becomes a near intractable problem.

The alternative, dimensionality reduction (DR), is advantageous because all of the original data is simply transformed from the original high dimensional feature space to a space of eigenvectors, capable of describing the data in far fewer dimensions. The largest eigenvectors represent the direction along which the greatest variability in the dataset occurs. Advantages of DR over feature selection include (i) representation of data structure in far fewer dimensions and (ii) the visualization of individual data classes and possibly subclasses within the high dimensional data.

The most popular method for DR is Principal Components Analysis (PCA). PCA finds orthogonal eigenvectors which account for the greatest amount of variability in the data. However, its basic intuitions lie under the assumption that the data is linear. These embedded eigenvectors represent low dimensional projections of linear relationships between data points in high dimensional space. Dai et al. [5] and Shi et al. [2] have independently tested the efficacy of PCA in improving the classification of gene expression datasets. Recently, methods such as Graph Embedding [6], Isometric mapping (Isomap) [7], and Locally Linear Embedding [8] have been developed to reduce the dimensionality of nonlinear data under the assumption that the underlying distribution is nonlinear. The structure of nonlinear data can be thought of as a high order curve or manifold where the geodesic distance between two points on the manifold is greater than their Euclidean distance would suggest. Nonlinear methods attempt to map data along this nonlinear manifold by assuming only neighboring points to be similar enough to be mapped linearly with minimal error. The nonlinear manifold can then be reconstructed based on these locally linear assumptions, providing the groundwork for a nonlinear mapping based on the true distances between any two data points. In general however, the choice of DR methods for the analysis of medical data has been relatively arbitrary. Although there is widespread evidence [2,9,10] to suggest that medical data such as genomic and proteomic expression studies are nonlinear, surprisingly few researchers have attempted nonlinear DR methods for this purpose. Shi and Chen [2] have evaluated the use of LLE in comparison with PCA for improving classification in leukemia, lymphoma and colon gene expression datasets. Dawson et al. [9] explored the utility of Isomap in comparison with PCA and linear multidimensional scaling (MDS) in oligonucleotide datasets, and Nilsson et al. [10] independently compared Isomap with MDS to reveal structures in microarray data related to biological phenomena. Madabhushi et al. [6] demonstrated the use of graph embedding to detect the presence of new tissue classes on high dimensional prostate MRI studies. While significant work in comparing classifier performance on cancer studies has been done [4,11], no serious quantitative comparisons involving multiple DR algorithms have been done in the context of maximizing classification accuracy.

The primary motivation of this paper is twofold. Firstly, by quantitatively comparing the performance of multiple linear and nonlinear DR methods, we can determine the appropriate technique to precede classification in high dimensional gene and protein expression studies. Secondly, we wish to demonstrate that nonlinear DR methods are superior compared to linear methods both from the perspective of classification and from the perspective of identifying and visualizing new classes within the data. In this work, we consider genomic and proteomic expression datasets from 7 separate studies corresponding to prostate, lung and ovarian cancers, as well as leukemia and lymphoma. Three different linear methods (PCA, linear discriminant analysis (LDA) [3], linear MDS [10]) and three nonlinear DR methods (graph embedding [6], Isomap [7], and LLE [8]) are applied to each of the datasets. The low dimensional embedding vectors, obtained from each DR method and for each dataset, are then supplied to a support vector machine classifier and a decision tree classifier. The accuracy of each classifier in distinguishing between cancer and non-cancer classes is thus used to gauge the efficacy of each of the DR methods. In addition to classification, we also quantitatively compare each of the DR methods in terms of their ability to detect new sub-classes within the data.

The organization of the rest of this paper is as follows. In Section 2, we will give a brief overview of the DR methods considered in this work. In Section 3 we describe our experimental design. Our qualitative and quantitative results in comparing the different DR methods are presented in Section 4. Finally, we present our concluding remarks in Section 5.

2 Description of Dimensionality Reduction Methods

In this section we briefly describe the 3 linear (PCA, MDS, LDA) and 3 nonlinear DR methods (Graph Embedding, Isomap, LLE) considered in this study.

2.1 Linear Dimensionality Reduction Methods

Principal Components Analysis (PCA): PCA has been widely documented as an effective means for analyzing high dimensional data [5]. Briefly, PCA applies a linear transformation to the data that allows the variance within the data to be expressed in terms of orthogonal eigenvectors. The eigenvectors that contain the most variance in the data represent the *principal components*.

Linear Discriminant Analysis (LDA): LDA [3] takes into account class labels to find intra-class correlations in the dataset. Assuming there is a linear hyperplane that can maximize separation between the two classes, LDA projects features that maximally account for this inter-class difference. While LDA has been useful as both a DR method and a classifier, it is limited in handling sparse data in which a Gaussian distribution of data points does not exist [3].

Classical Multidimensional Scaling (MDS): MDS [10] is implemented as a linear method that uses Euclidean distances between each pair of points as a basis for a low dimensional data arrangement. From these input distances, MDS finds optimal positions for the data points in an arbitrary d-dimensional space by minimizing least square error. Thus, the relative Euclidean distances between points in low-dimensional embedding space are preserved. Note that classical MDS differs from nonlinear variants of MDS such as nonmetric MDS, which do not preserve input Euclidean distances.

2.2 Nonlinear Dimensionality Reduction Methods

Graph Embedding (GE): The GE algorithm [6] performs a series of normalized cuts on the data to partition it into clusters of data points. These cuts are made where minimal similarities exist (decided using a similarity matrix of pairwise Euclidean distances). In this manner, similarity can be discerned by inter- and intra-cluster distances, where points within a cluster are deemed similar and points belonging to separate clusters are deemed dissimilar. Separating points by GE allows for the separation of objects within complex nonlinear structures, where objects cannot otherwise be discriminated by linear DR methods.

Isometric Mapping (Isomap (ISO)): The Isomap algorithm [7] is essentially one that optimizes classical MDS for the nonlinear case. Isomap finds the nonlinear manifold on which the data is expected to lie through the use of a neighborhood map, which assumes linearity only between its **k** nearest neighbors defined by the user. By connecting each point only to its nearest neighbors, a path representing the geodesic distances between two points can be approximated by finding the shortest path through the neighborhood mapping. These new geodesic distances represent the true distances between points and serve as input into classical MDS, where a more accurate low dimensional representation can be constructed.

Locally Linear Embedding (LLE): LLE [8] attempts to create a low dimensional representation of the global structure through the preservation of the local structure by assuming only nearby points to be linear. Local linearity is achieved by weighting only the **k**-nearest neighbors of each data point. A total of d new embedding vectors are then reconstructed by these linear weights and by minimizing the embedding cost function in the new d-dimensional coordinate system.

3 Experimental Design

In this Section, we first briefly describe the datasets considered in this study along with a description of the parameter settings for the DR methods (Section 3.1), followed by a brief description of the classifiers considered (Section 3.2) and our model for performing a quantitative comparison of the different DR methods (Section 3.3).

3.1 Description of Datasets and Parameter Settings

To evaluate the different DR methods, we chose 7 publicly available datasets corresponding to high dimensional protein and gene expression studies[1]. The size of the datasets ranged from 34 to 253 patient samples and comprised from between 4026 to 15154 genes. Table 1 lists all the datasets on which we tested our DR methods. Note that for each dataset considered, the number of samples is significantly smaller than the dimensionality of the feature space. For the ALL-AML Leukemia dataset, two classes were considered: Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). For the DLBCL-Harvard dataset, the 2 classes considered were Diffuse large B-cell Lymphoma (DLBCL) and Follicular Lymphoma (FL). Lastly, the Lung Cancer dataset contains two types of lung cancer (mesothelioma (MPM) and adenocarcinoma (ADCA)).

Table 1. Gene expression and proteomic spectra datasets considered in this study

Dataset	*Samples*	*Genes*	*Class Description*	*Source*
(1) ALL-AML Leukemia	34	7129	20 ALL, 14 AML	Golub et al. [12]
(2) DLBCL-Harvard	77	6817	58 DLBCL,19 FL	Shipp et al. [13]
(3) Lung Cancer	148	12533	15 MPM, 134 ADCA	Gordon et al. [14]
(4) Lung Cancer-Michigan	96	7129	86 Tumor, 10 Normal	Beer et al. [15]
(5) Ovarian Cancer	253	15154	162 Tumor, 91 Normal	Petricoin et al. [16]
(6) Prostate Cancer	34	12600	25 Tumor, 9 Normal	Singh et al. [17]
(7) Types of Diffuse Large B-cell Lymphoma	47	4026	24 Germinal, 23 Activated	Alizadeh et al. [18]

For each of the datasets D_j, $1 \leq j \leq 7$, we applied each of 6 DR methods M, where $M \in \{PCA, LDA, MDS, GE, ISO, LLE\}$. For each method M and dataset D_j, we obtained a set $S^d_{D_j,M} \geq \left\{E^1_{D_j,M}, E^2_{D_j,M}, ..., E^d_{D_j,M}\right\}$ of d dominant eigenvectors. The number of principal eigenvectors d used to classify the objects $c \in D_j$ were varied from 2 to 8 in order to find the optimal d-dimensional space in which the 2 classes were most easily separable.

[1] The 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

3.2 Classifiers

To perform our classification, we input a set $S^d_{D_j,M}$ of eigenvectors to the following 2 machine learning classifier methods: Support Vector Machines (SVMs) and C4.5 Decision Trees. Both require the use of a training set to construct a prediction model for new data. SVMs project the input data to a higher dimensional space to find a hyperplane that gives the greatest separation between the data classes. This hyperplane along with 2 parallel support vectors serve as a boundary in which a prediction can be made for new data. Decision Trees create a predictor wherein new samples are categorized based on several conditional statements. For each condition, the algorithm associates a certain likelihood that a sample falls into a particular category and refines the class hypothesis before a final decision is made.

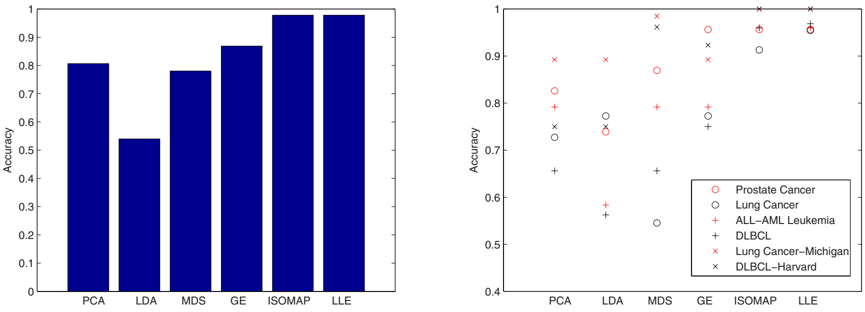
Since the classifiers were being used to evaluate the DR methods' ability to separate 2 classes, a simple linear kernel was chosen for SVM. The linear kernel draws a d-dimensional hyperplane to act as a decision boundary between the two separated classes. To train the 2 classifiers, we set aside 1/3 of the samples in each dataset $D_j, 1 \leq j \leq 7$, for 3-fold cross validation. Using the best samples from cross validation, we determined the parameter settings for both classifiers. After the model parameters were learned, the same parameter values for SVM and C4.5 were used to test the remaining 2/3 objects in each D_j.

3.3 Quantitative Evaluation of DR Methods

The accuracy of the SVM and C4.5 classifiers on 7 datasets $D_j, 1 \leq j \leq 7$ was quantitatively evaluated using the class labels provided in the gene expression studies. We define accuracy as the ratio of the number of objects $c \in D_j, 1 \leq j \leq 7$, correctly labeled by the classifier to the total number of tested objects in each D_j. We denote the classification accuracy of SVMs on dataset $S^d_{D_j,M}$ by SVM($S^d_{D_j,M}$) and the corresponding accuracy of the C4.5 Decision Trees by C4.5($S^d_{D_j,M}$). To determine whether the classifier results from the nonlinear and linear DR methods were significantly different, we performed a paired student t-test wherein we compared SVM($S^d_{D_j,M}$) for $M \in \{PCA, LDA, MDS\}$ versus SVM($S^d_{D_j,M}$) for $M \in \{GE, ISO, LLE\}$ across all D_j. The t-test was similarly repeated for C4.5($S^d_{D_j,M}$). The difference between SVM($S^d_{D_j,M}$) or C4.5($S^d_{D_j,M}$) for each pair of methods (1 linear and 1 nonlinear) was deemed to be statistically significant if $p \leq 0.05$. The linear and nonlinear DR methods were also semi-quantitatively compared (i) using 2-D embedding plots to evaluate their ability to distinguish between the cancer and non-cancer clusters and (ii) to potentially identify and visualize the presence of new sub-classes. In order to visualize the embedding of the data in the low dimensional space obtained via the DR methods, we plotted the dominant eigenvectors obtained for each M for each object $c \in D_j$ against each other (e.g. $E^2_{D_j,M}$ versus $E^3_{D_j,M}$).

4 Results and Discussion

In Section 4.1 we present the results of quantitative comparison of the different DR methods in terms of their classification accuracy obtained from the SVM and C4.5 classifiers. In Section 4.2 we present quantitative graphical plots comparing the ability of the DR methods to separate the data classes and also in identifying and visualizing the presence of new classes.



(a) (b)

Fig. 1. (a) Average C4.5($S^d_{D_j,M}$) for $d = 6$ and (b) SVM($S^d_{D_j,M}$) for $d = 5$ for each of 6 datasets following DR by *PCA*, *LDA*, *MDS*, *GE*, *ISO*, and *LLE*. Note that for both classifiers, the nonlinear methods consistently outperform the linear methods.

4.1 Quantitative Evaluation of DR Methods Via Classifier Accuracy

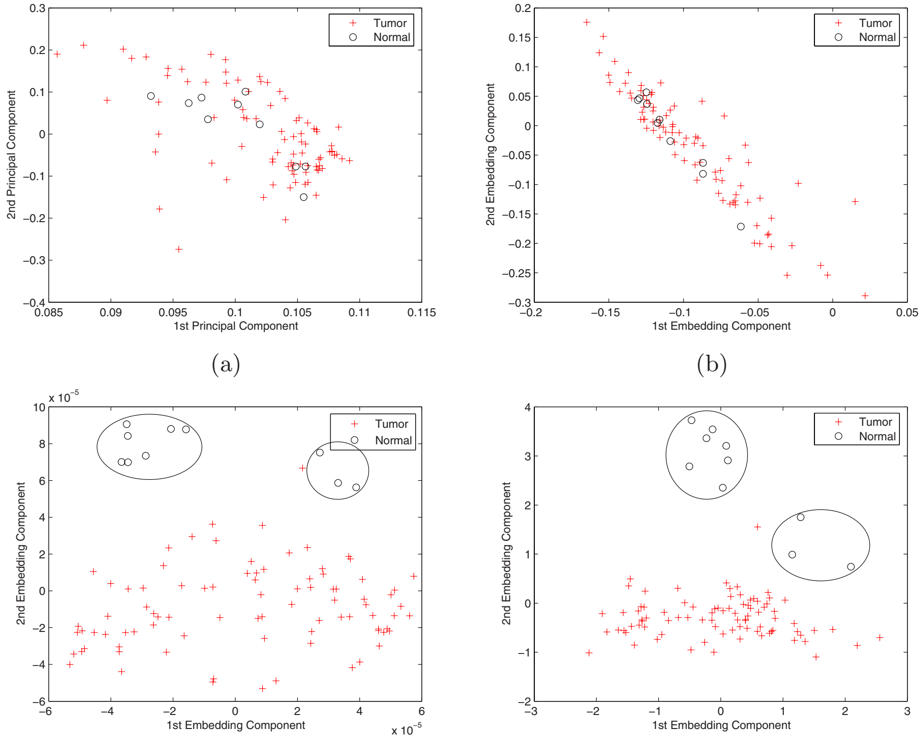
In Figure 1(a), we show the average accuracy results obtained with the C4.5 classifier, C4.5($S^d_{D_j,M}$) for $1 \leq j \leq 6$. We obtained our best results for $d = 5$ for SVMs and $d = 6$ for C4.5 Decision Trees. From Figure 1(a), it is clear that embeddings from nonlinear DR methods (*GE*, *ISO*, *LLE*) lead to better overall accuracy than with linear DR methods (*PCA*, *LDA*, *MDS*). Isomap and LLE overall were the most accurate while LDA performed the worst. In

Table 2. C4.5($S^d_{D_j,M}$) for each of 7 datasets following DR by *PCA*, *LDA*, *MDS*, *GE*, *ISO*, and *LLE* for $d = 6$

Dataset	*PCA*	*LDA*	*MDS*	*GE*	*ISO*	*LLE*
(1) ALL-AML Leukemia	62.5	41.7	62.5	91.7	95.0	95.0
(2) DLBCL-Harvard	69.2	40.4	84.6	86.5	96.9	96.9
(3) Lung Cancer	67.7	84.6	70.8	98.5	100.0	100.0
(4) Lung Cancer-Michigan	67.7	84.6	69.2	98.5	100.0	100.0
(5) Ovarian Tumor	55.6	59.2	61.5	59.2	59.8	63.3
(6) Prostate Cancer	100.0	47.8	87.0	82.6	100.0	100.0
(7) Types of Diffuse Large B-cell Lymphoma	93.8	59.4	90.6	93.8	95.0	95.0

Table 3. p-values obtained by a paired student t-test of SVM($S^d_{D_j,M}$) across 7 data dimensions $d \in \{2, ...8\}$ comparing linear versus nonlinear DR methods for $1 \leq j \leq 7$. Note that the numbers listed in the first column refer to the datasets given in Table 1.

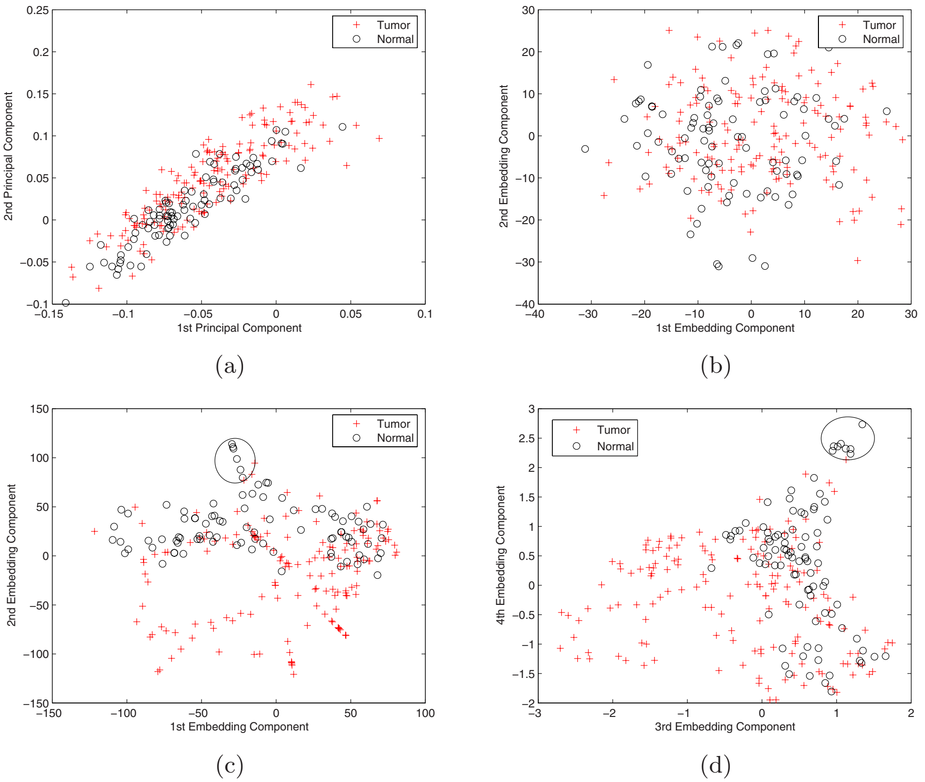
	GE	*GE*	*GE*	*ISO*	*ISO*	*ISO*	*LLE*	*LLE*	*LLE*
Dataset	*vs*	*vs*	*vs*	*vs*	*vs*	*vs*	*vs*	*vs*	*vs*
	PCA	*LDA*	*MDS*	*PCA*	*LDA*	*MDS*	*PCA*	*LDA*	*MDS*
(1)	.068	$8x10^{-5}$	.057	$7x10^{-5}$	$6x10^{-8}$	.002	$7x10^{-5}$	$6x10^{-8}$	.002
(2)	.373	$3x10^{-4}$	.925	.012	$6x10^{-5}$	.014	.012	$6x10^{-5}$	.014
(3)	.361	.852	.691	.003	.009	$4x10^{-4}$	.002	.006	$4x10^{-4}$
(4)	.706	.063	1.000	.004	$3x10^{-8}$	.015	.005	$2x10^{-16}$	.011
(5)	.478	.063	.412	.008	.004	.003	.002	.003	$2x10^{-4}$
(6)	.156	.001	.045	$2x10^{-5}$	$4x10^{-8}$	.012	$8x10^{-5}$	$3x10^{-8}$	.019
(7)	.005	10^{-4}	.074	$8x10^{-5}$	$7x10^{-6}$	.001	$8x10^{-5}$	$7x10^{-6}$	.001



(a) (b)

(c) (d)

Fig. 2. Embedding plots were obtained by graphing the 2 dominant eigenvectors against each other for (a) PCA, (b) LDA, (c) GE, and (d) LLE for the Lung Cancer-Michigan dataset. Note that while linear methods, PCA and LDA, are unable to distinguish between the 2 classes, the nonlinear methods, GE and LLE, are able to not only clearly distinguish between the 2 groups but also permit visualization of 2 possible normal class sub-clusters (indicated by superposed ellipses in (c) and (d)).



(a)

(b)

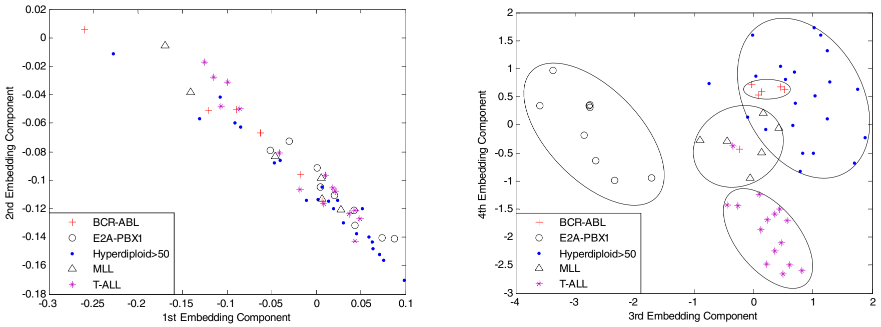
(c)

(d)

Fig. 3. Embedding plots were obtained by graphing the 2 dominant eigenvectors against each other for (a) PCA, (b) MDS, (c) Isomap, and (d) LLE for the Ovarian Cancer dataset. As in Figure 3, we can appreciate that nonlinear methods, Isomap and LLE, are able to distinguish between the 2 classes and also permit visualization of a possible normal class sub-cluster (indicated by superposed ellipses in (c) and (d)).

Table 2, we show C4.5($S^d_{D_j,M}$), for $1 \leq j \leq 7$, for all 6 DR methods, for $d = 6$. Our results indicate an improvement in accuracy for the nonlinear DR over linear DR methods. In Table 3 are listed p-values for the paired student t-tests obtained for SVM($S^d_{D_j,M}$) across 7 data dimensions ($d \in \{2, ..., 8\}$) for each paired comparison of a linear and non-linear DR method. Hence we compared the following pairs of methods: PCA/GE, LDA/GE, MDS/GE, PCA/Isomap, LDA/Isomap, MDS/Isomap, PCA/LLE, LDA/LLE, MDS/LLE for each of the 7 datasets considered. As the results in Table 3 indicate, differences in classification accuracy for pairs PCA/GE and MDS/GE were not statistically significant ($p \geq 0.05$) while all corresponding paired comparisons involving LLE and Isomap were statistically significantly more accurate compared to linear DR methods.

The results in Figure 1 and Tables 2 and 3 clearly suggest that nonlinear DR methods result in higher statistically significant accuracy compared to linear DR methods, as determined by 2 separate classifiers. Additionally, Isomap and



(a) (b)

Fig. 4. Embedding plots were obtained by graphing 2 dominant eigenvectors against each other for 5 different types of Acute Lymphoblastic Leukemia. In Figure 4(a), an embedding plot for linear LDA is compared with Figure 4(b), an embedding plot of nonlinear LLE. Note that the linear method fails to distinguish the ALL sub-classes in the reduced eigenspace, while the LLE plot clearly reveals the presence of 5 distinct clusters (indicated by superposed ellipses).

LLE were found to generate the most useful embeddings resulting in highest classification accuracy.

4.2 Semi-quantitative Evaluation of Dimensionality Reduction Methods Via Class Separation and Novel Class Detection

We evaluated the efficacy of nonlinear DR methods in identifying new data classes and intermediate cancer types. This was done by visual inspection of 2-D cluster plots obtained by plotting $E^1_{D_j,M}$ versus $E^2_{D_j,M}$ for each of the 6 DR methods. The results of the 2-D embedding plots for the Lung Cancer-Michigan and Ovarian Cancer datasets are shown in Figures 2 and 3 respectively. In Figure 2, two distinct sub-classes can be distinguished in the normal class (indicated by superposed ellipses) for GE (Figure 2(c)) and LLE (Figure 2(d)) as well as a clear, distinct separation between the cancer and non-cancer classes. For the linear embeddings obtained via PCA and LDA (shown in Figures 2(a) and (b) respectively), there appears to be significant overlap between cancer and non-cancer classes. The poor class separation is reflected in the poor classification accuracy obtained with both the SVM and C4.5 Decision Tree classifiers in Figure 1 and Table 2.

In Figure 3, we have shown the comparison of embedding plots for linear PCA (Figure 3(a)) and MDS (Figure 3(b)) against nonlinear methods, Isomap (Figure 3(c)) and LLE (Figure 3(d)), on the Ovarian Cancer dataset. In Figures 3(c) and (d), one can appreciate a sub-cluster of normal samples (indicated by superposed ellipses), possibly suggesting pre-malignant cases. More importantly, the fact that this unusual clustering is present in both nonlinear DR algorithms strongly suggests the validity of the identified sub-clusters and the utility of nonlinear DR methods in visualizing biological relationships between samples and in new class discovery.

Since we were unable to quantitatively evaluate the validity of the sub-clusters detected by the nonlinear DR methods in Figures 2 and 3, we also compared linear and nonlinear methods on a multiclass dataset. Our aim is to demonstrate that nonlinear DR methods are more capable of detecting subtypes of Acute Lymphoblastic Leukemia (ALL) [19]. Figures 4(a) and (b) show plots comparing LDA and LLE in clustering 5 subtypes of ALL. As shown in 4(a), LDA does not provide any inter-class distinction while the embedding provided by LLE in 4(b) enables easy separation between the multiple sub-classes in ALL.

5 Concluding Remarks

In this paper we presented the results of quantitatively comparing the performance of 6 different DR methods (3 linear, 3 nonlinear) from the perspective of classification and the identification of new object classes in high dimensional gene and protein expression datasets for prostate, lung, and ovarian cancers, as well as for leukemias and lymphomas. The eigenvectors obtained from each of the different DR methods were supplied to two different classifiers (SVMs and Decision Trees) to distinguish between data classes within 7 different gene and protein expression studies. Classification accuracy with both SVMs and C4.5 Decision Trees were found to be consistently higher when using features obtained by nonlinear DR methods compared to linear methods. Among the nonlinear methods, LLE gave the highest overall accuracy. In addition to distinguishing between known classes, we were also able to identify the presence of several potential sub-clusters via nonlinear DR techniques. For most datasets, all the nonlinear DR methods outperformed the corresponding linear methods, differences being statistically significant in most cases. In future work we intend to quantitatively evaluate the validity of our results on several addition datasets.

Acknowledgments. This work was made possible due to grants from the Coulter foundation (WHCF 4-29349, WHCF 4-29368), the Busch Biomedical Award and the Technology Commercialization Fund at Rutgers University.

References

1. Peng Y. A novel ensemble machine learning for robust microarray data classification. Comput Biol Med. 2006, vol.36[6], pp.553-73.
2. Shi C. and Chen L. Feature Dimension Reduction for Microarray Data Analysis Using Locally Linear Embedding. APBC 2005. pp.211-217.
3. Ye J et al. Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data. IEEE/ACM Trans. Comput. Biology Bioinform. 2004, vol.1[6], pp.181-190.
4. Tan AC and Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics. 2003, pp.65-83.

5. Dai J et al. Dimension Reduction for Classification with Gene Expression Microarray Data. Statistical Applications in Genetics and Mol Biol. 2006, vol.5[1], pp.1-15
6. Madabhushi A et al. Graph Embedding to Improve Supervised Classification and Novel Class Detection: Application to Prostate Cancer. MICCAI 2005. pp.729-737.
7. Tenenbaum JB et al. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science. 2000, vol.290, pp.2319-2322.
8. Roweis ST and Saul LK. Nonlinear Dimensionality Reduction by Local Linear Embedding. Science. 2000, vol.290, pp.2323-2326.
9. Dawson K et al. Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. BMC Bioinformatics. 2005, vol.6, pp.195.
10. Nilsson J et al. Approximate geodesic distances reveal biologically relevant structures in microarray data. Bioinformatics. 2004, vol.20, pp.874-880.
11. Madabhushi A et al. Comparing Classification Performance of Feature Ensembles: Detecting Prostate Cancer from High Resolution MRI, Computer Vision Methods in Medical Image Analysis (In conjunction with ECCV). 2006, LNCS 4241, pp. 25-36.
12. Golub TR et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999, vol.286, pp.531-537.
13. Shipp MA et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002, vol.8, pp. 68-74.
14. Gordon GJ et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. 2002, vol.62, pp.4963-4967.
15. Beer D et al. Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. Nature Medicine. 2002, vol.8[8], pp.816-823.
16. Petricoin EF et al. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet. 2002, vol.359[9306], pp.572-577
17. Singh D et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002, vol.1, pp.203-209.
18. Alizadeh AA et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000, vol.403, pp.503-511.
19. Yeoh EJ, et al. Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling. Cancer Cell. 2002, vol.1[2], pp.133-143.